

Lithography, SoC Design, and Cost

Greg Yeric
ARM R&D

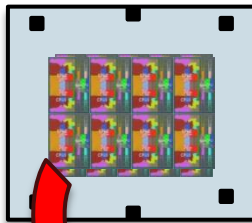
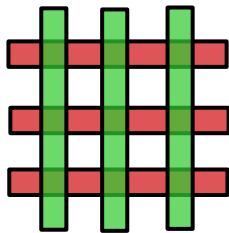
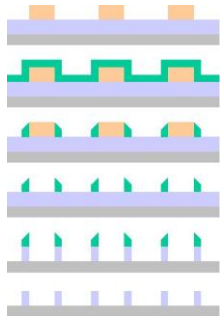


Cost Chain

**Designers and Lithographers bound
Together in need to support continued Moore's
Law reduction in cost/function**



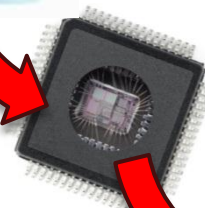
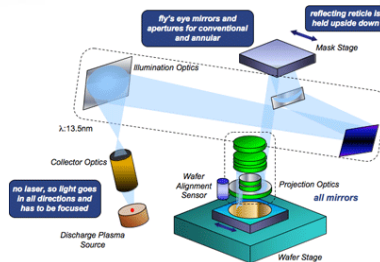
Lithography Cost



Try to shrink pitch
Without increasing wafer cost too much

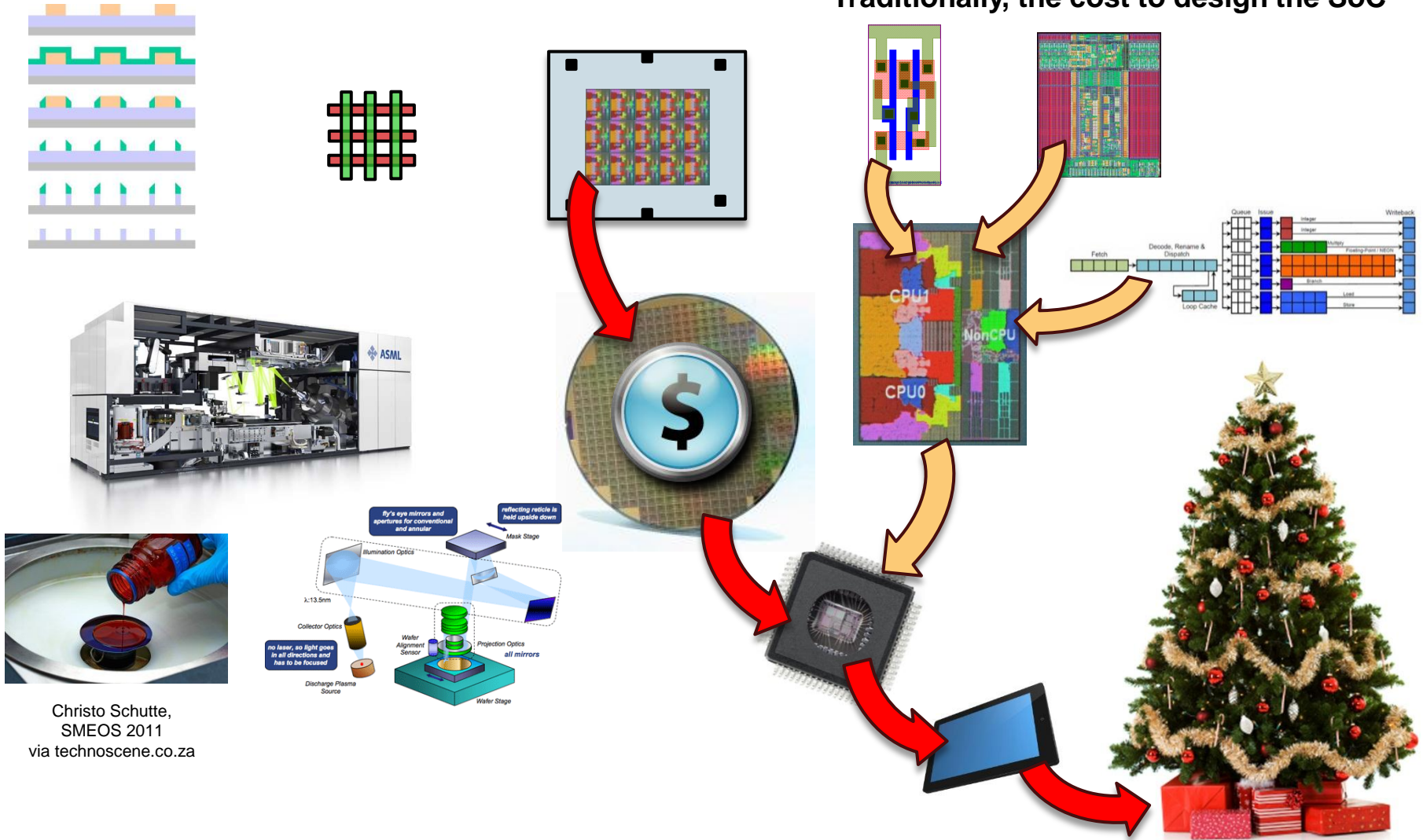


Christo Schutte,
SMEOS 2011
via technoscene.co.za



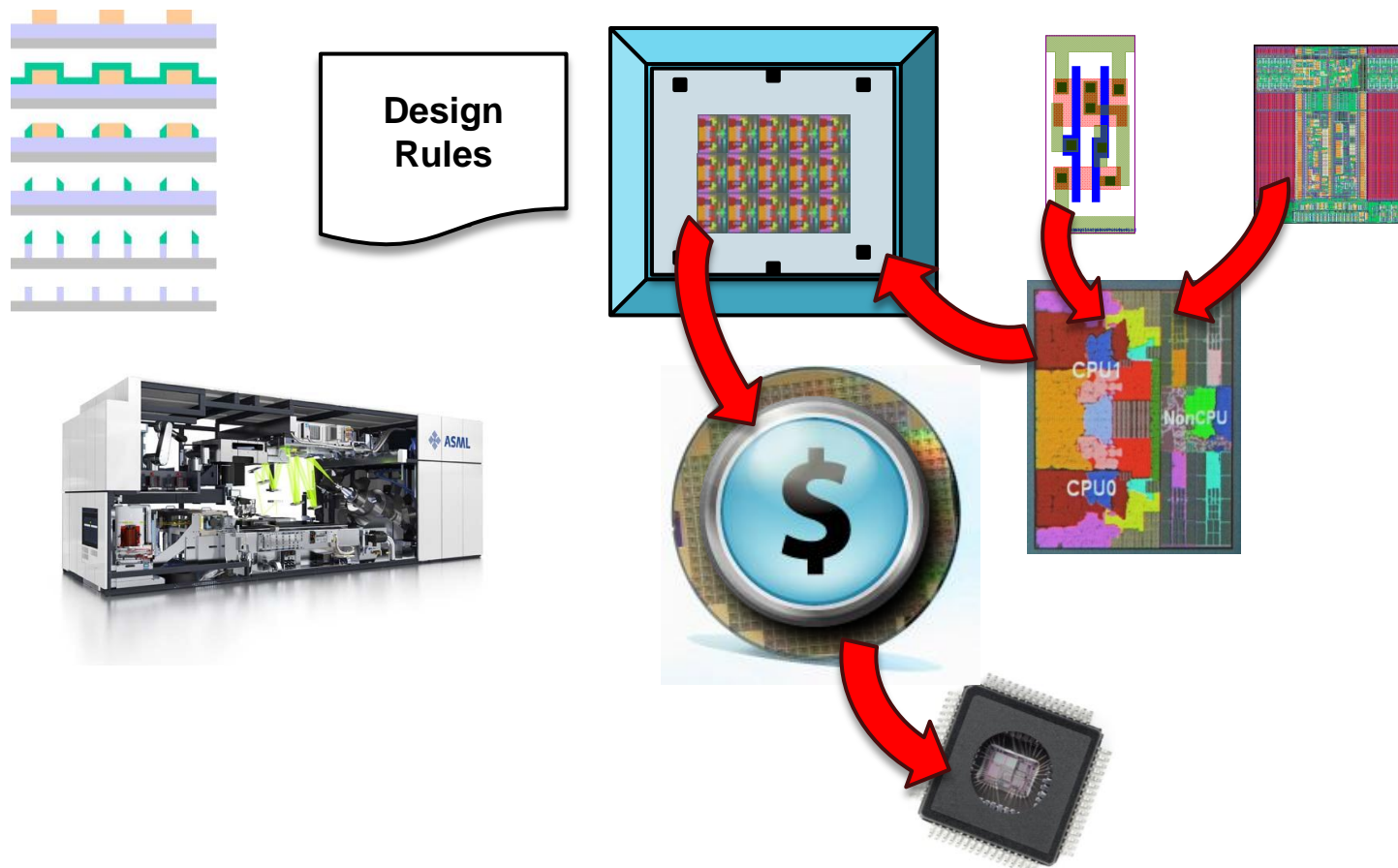
Design Cost

Traditionally, the cost to design the SoC



Christo Schutte,
SMEOS 2011
via technoscene.co.za

Design and Lithography Interactions



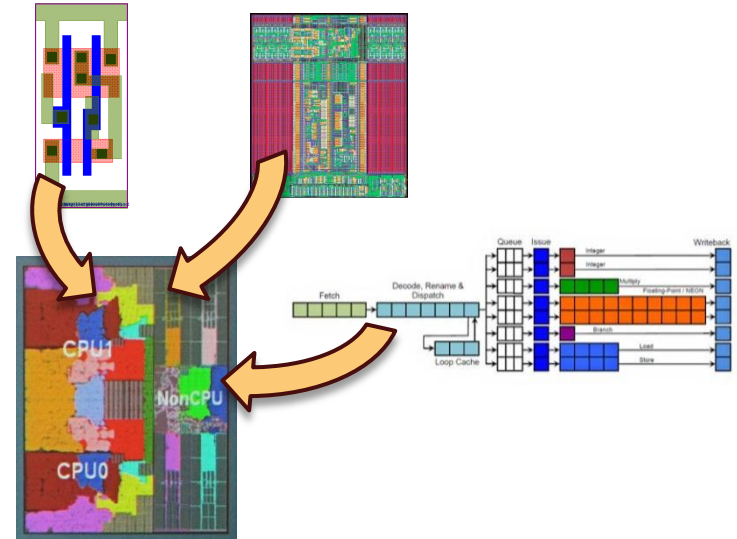
However, scope of this talk will examine ways in which cost enters the SoC through the interactions of lithography and design

Outline

- | | | |
|----|--|--------|
| 1. | Introduction and scope
(you are here) | 5 min |
| 2. | SoC Design Overview | 10 min |
| 3. | Lithography-Design Interactions and Cost | 25 min |

SoC Contents

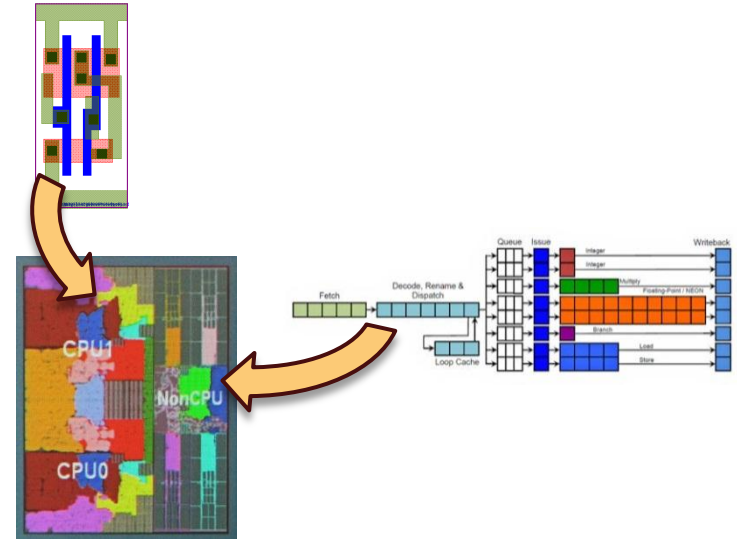
~ 20% I/O
~ 40% logic
~ 40% memory



I/O scaling not typically bound by litho constraints, so won't talk about that
40% memory is a bit high for most SoC's but a convenient number for this talk

SoC Contents

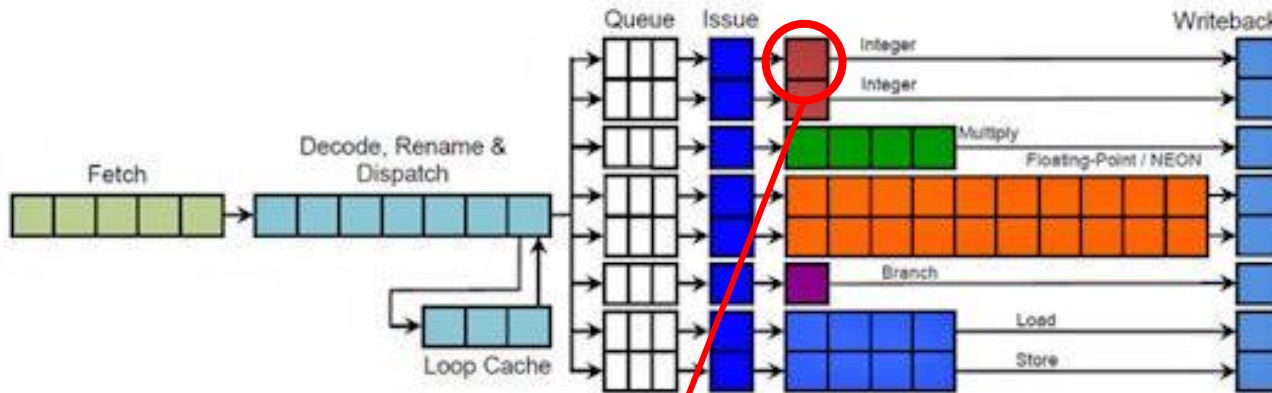
~ 20% I/O
~ **40% logic**
~ 40% memory



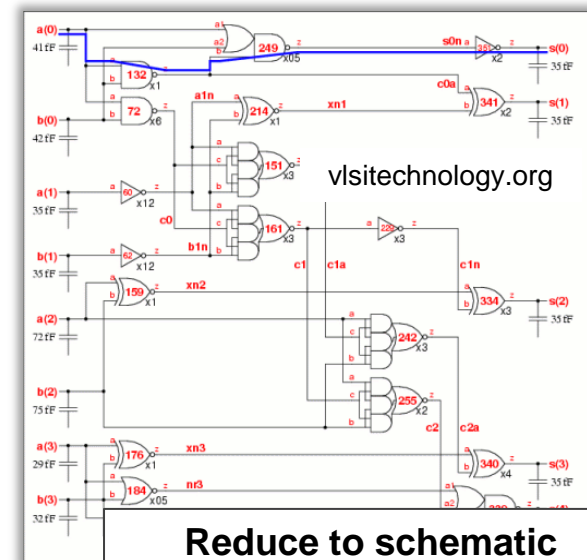
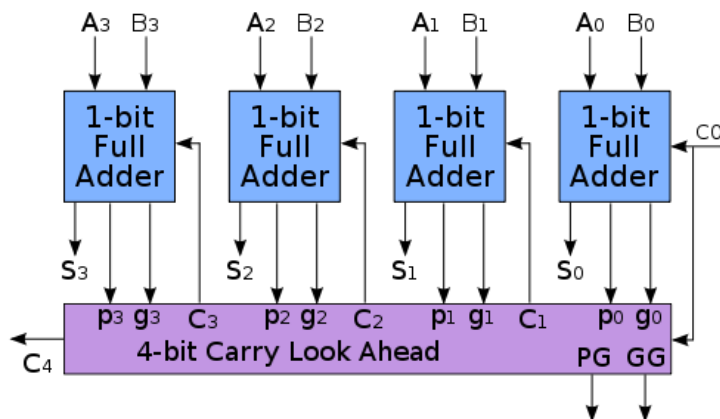
Synthesis, Place, and Route

To understand logic is to understand SP&R

Logic Flow: Synthesis



4-bit (full) adder

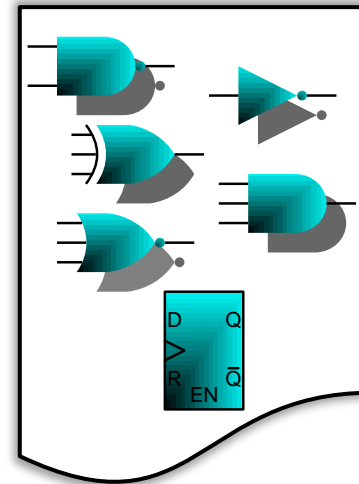
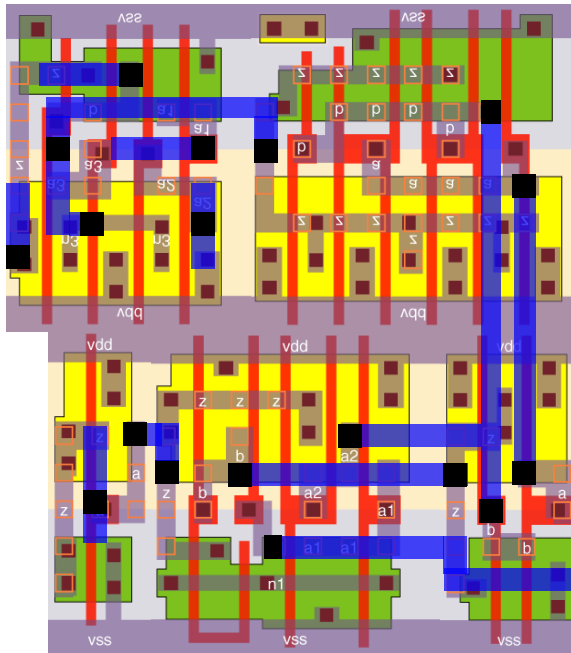


Reduce to schematic
Containing “standard cells”

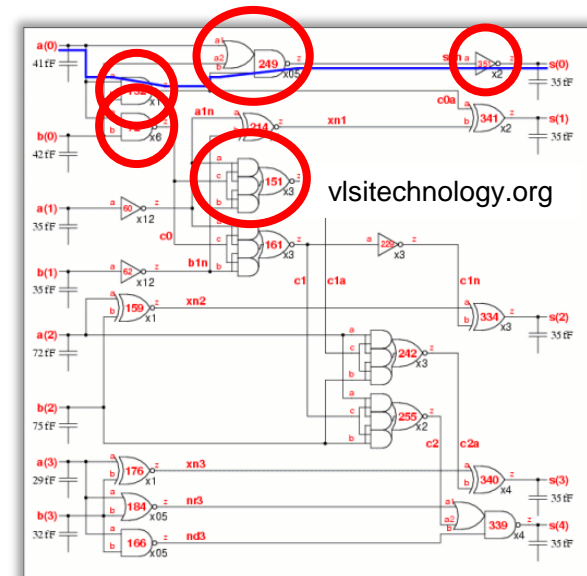
Logic Flow: Place and Route

Required Placements:

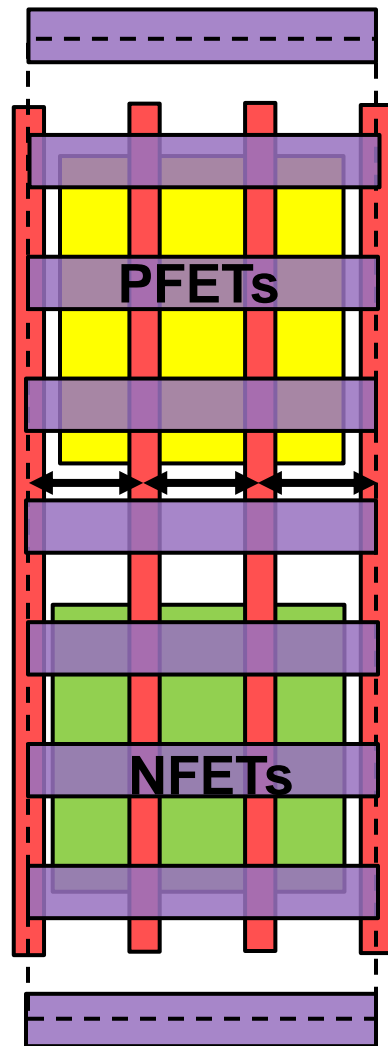
- Vertical and Horizontal Mirroring
- Placement at any gate pitch



**Standard
Cell
Library**



Standard Cell Lingo



VDD power rail

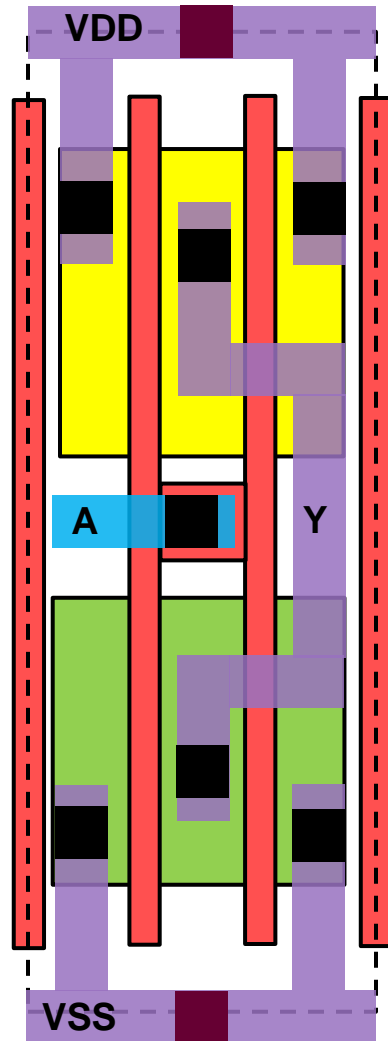
Minimum Metal Pitch = 1 “Track”

This cell is 8 tracks tall (8T)
(the key “standard” in standard cells)

This cell is 3 Contacted Poly Pitches Wide (3 CPP)

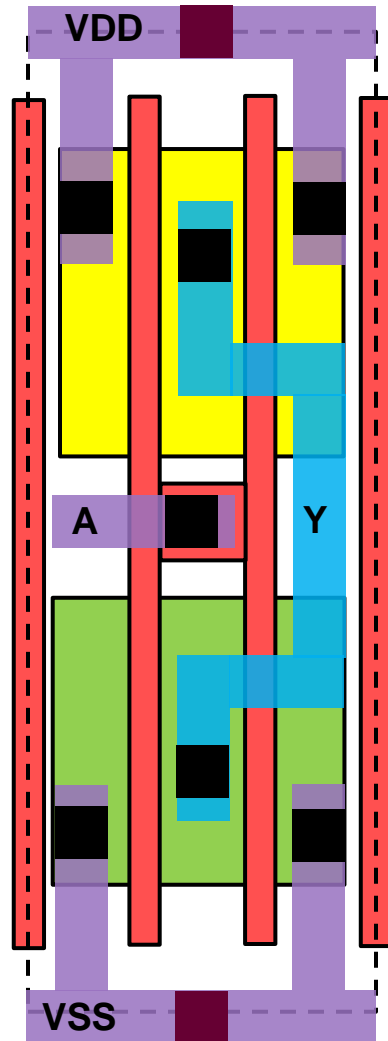
VSS power rail

Standard Cell Lingo



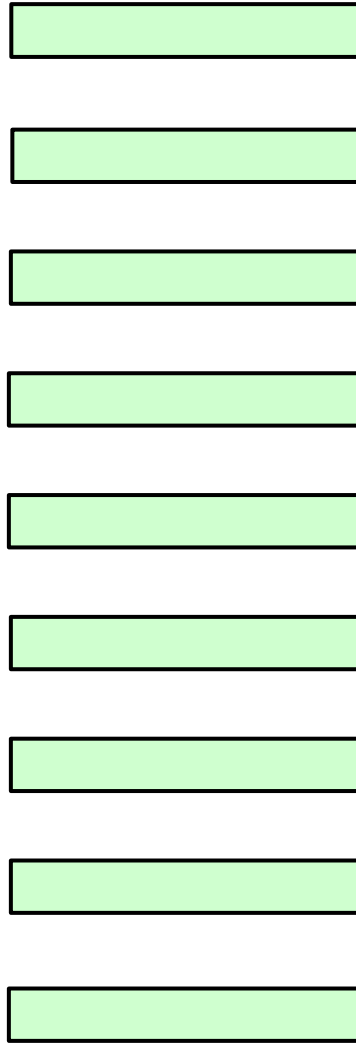
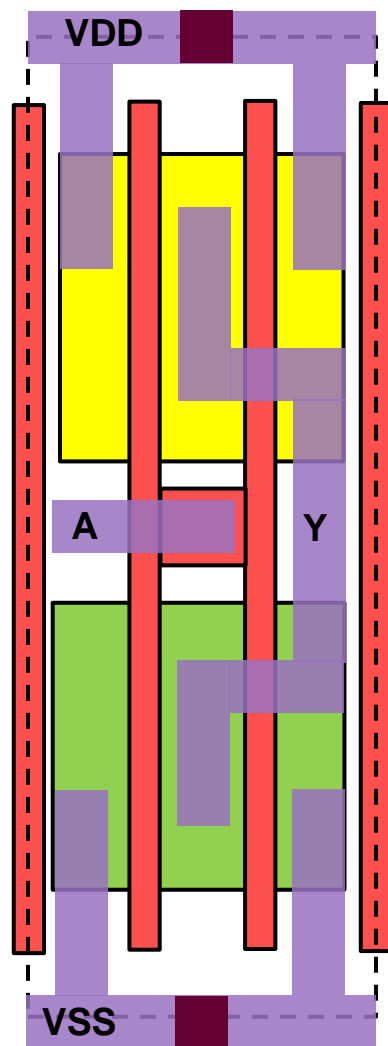
Input Pin

Standard Cell Lingo



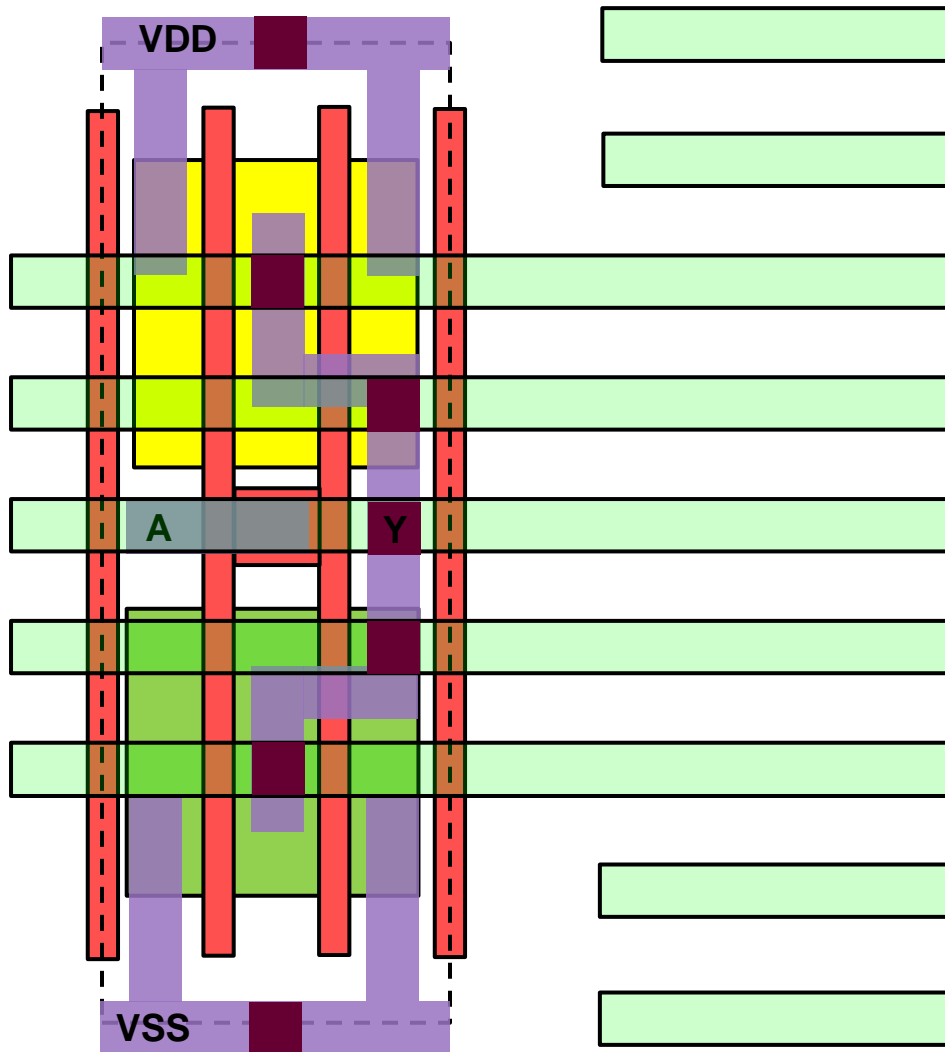
Output Pin

Standard Cell Lingo



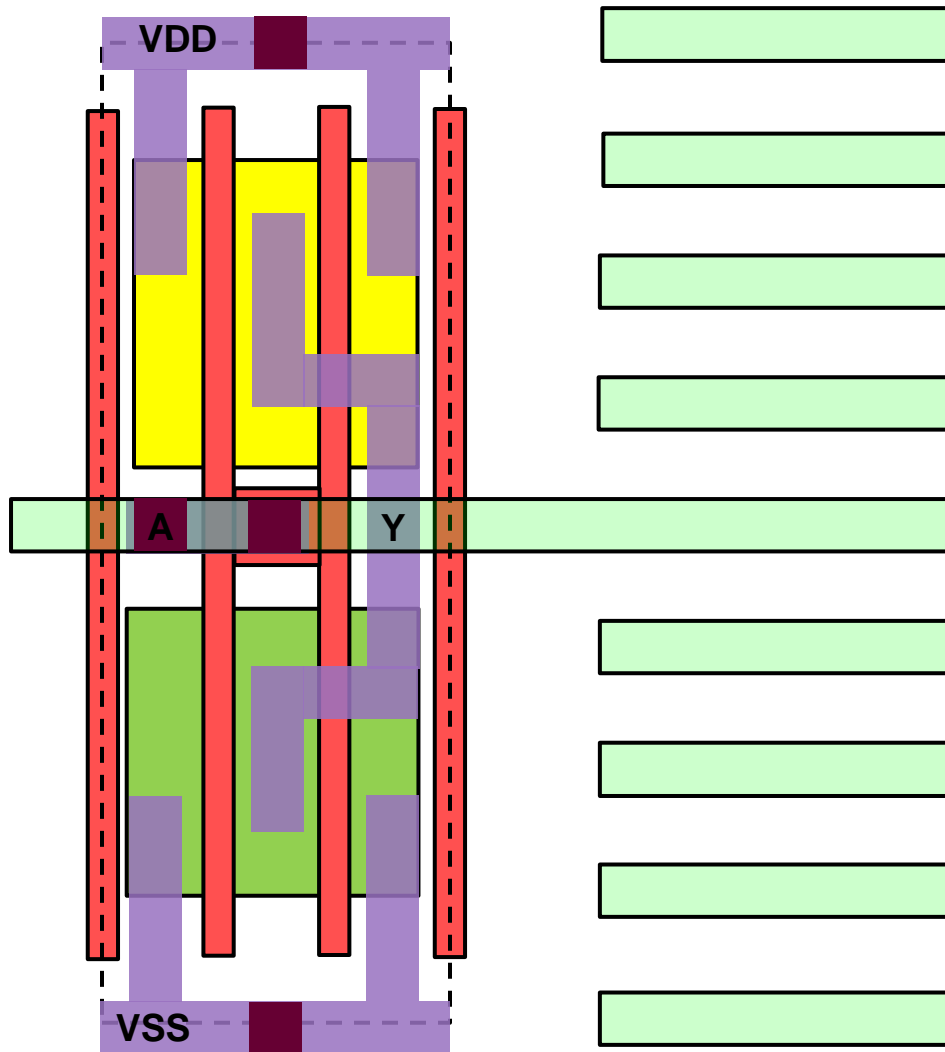
Metal 2 router tracks

Standard Cell Lingo



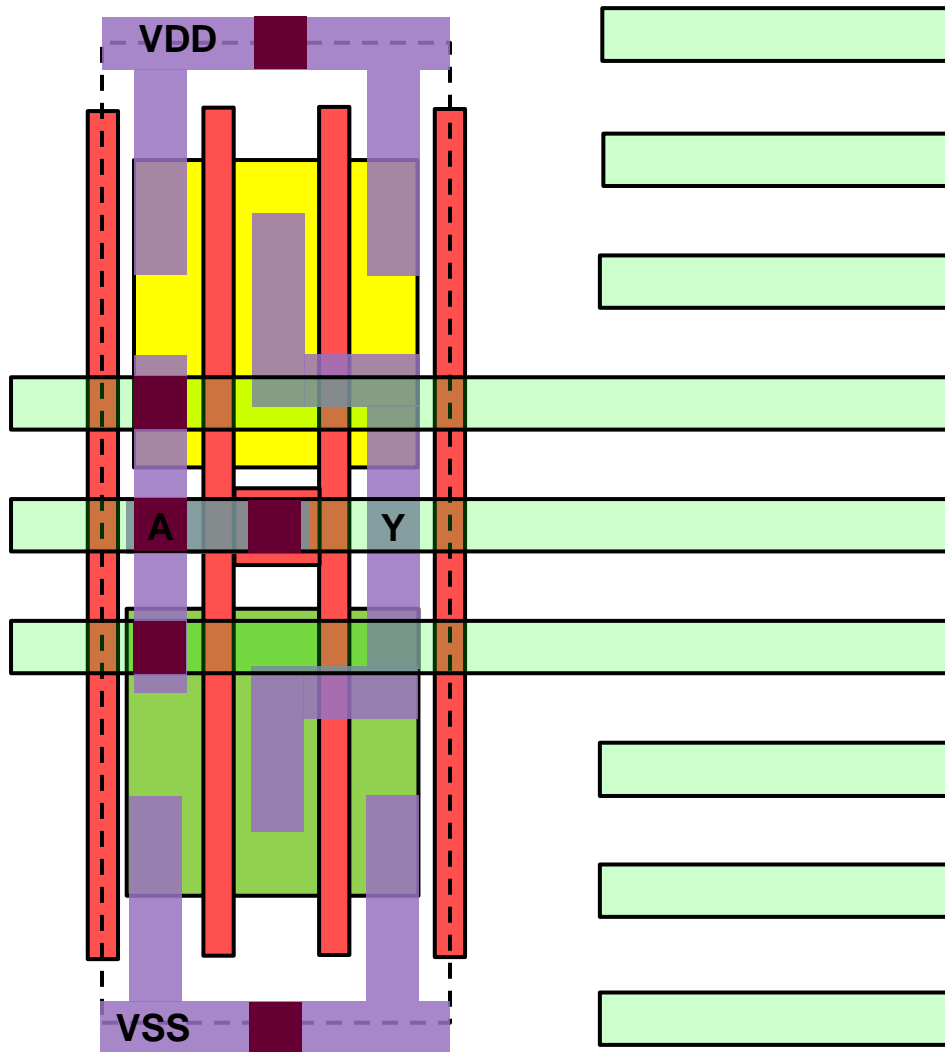
**There are 5 router
“hit points” on the Y pin**

Standard Cell Lingo



**There are 2 hit points
on the A pin...**

Standard Cell Lingo



...or 4 hit points,
If you get to count
“virtual” hit points

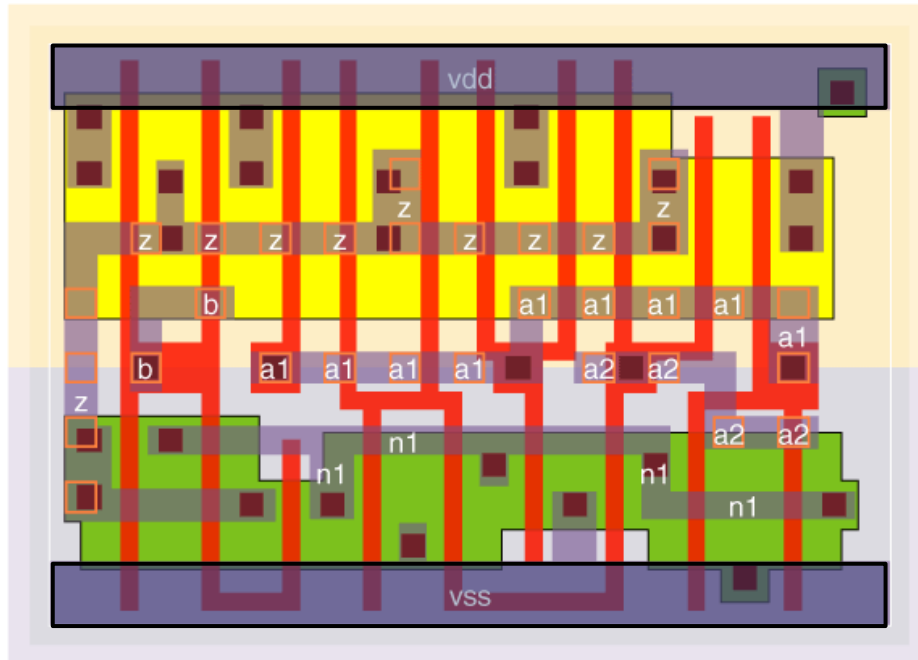
Rule of Thumb:

3 hit points: OK

2 hit points: Only in Moderation

1 hit point: Disaster

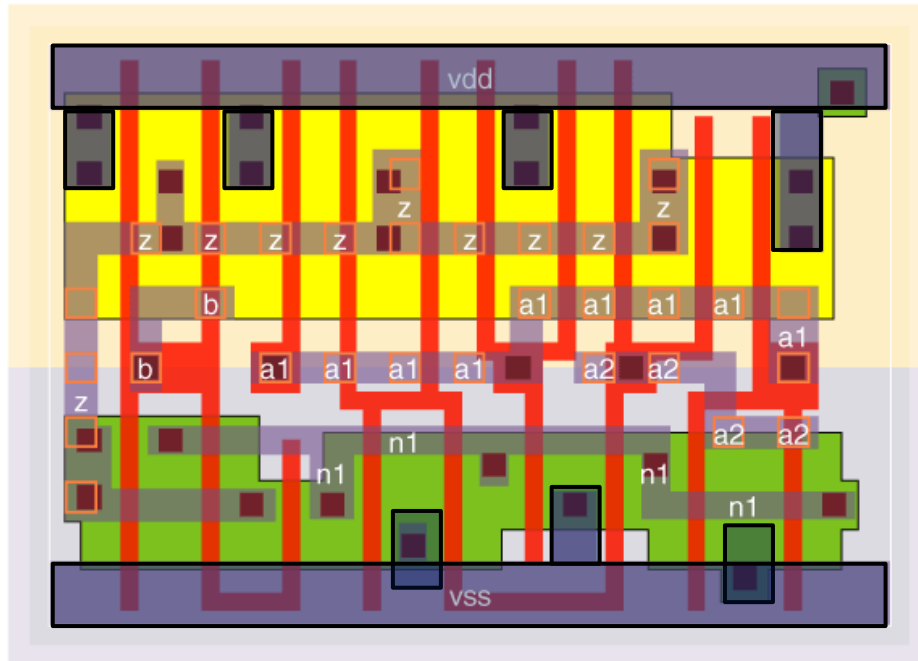
Canonical Standard Cell M1



Power rails are horizontal

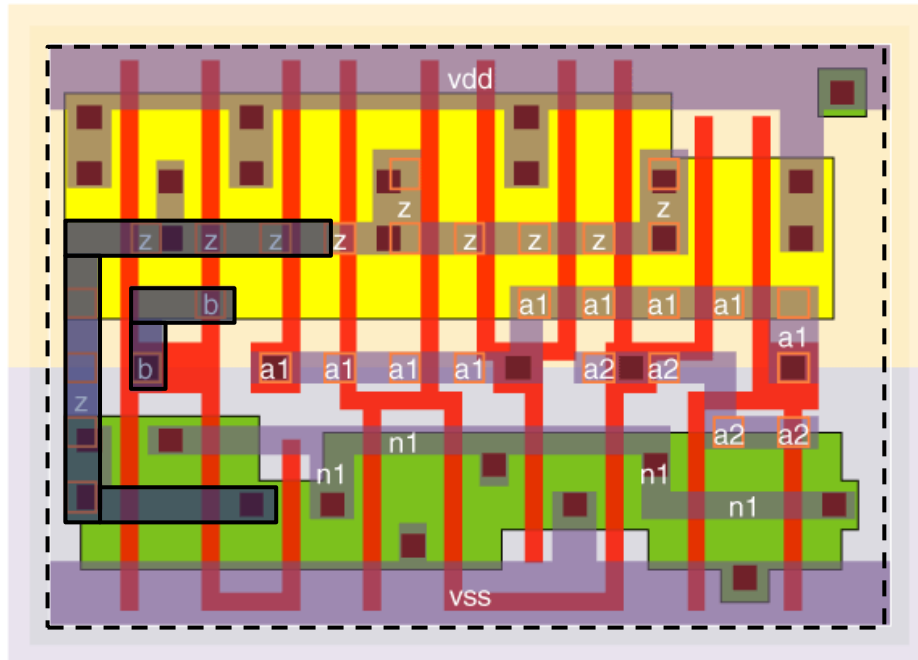
Power rails are (increasingly)
> minimum width

Canonical Standard Cell M1



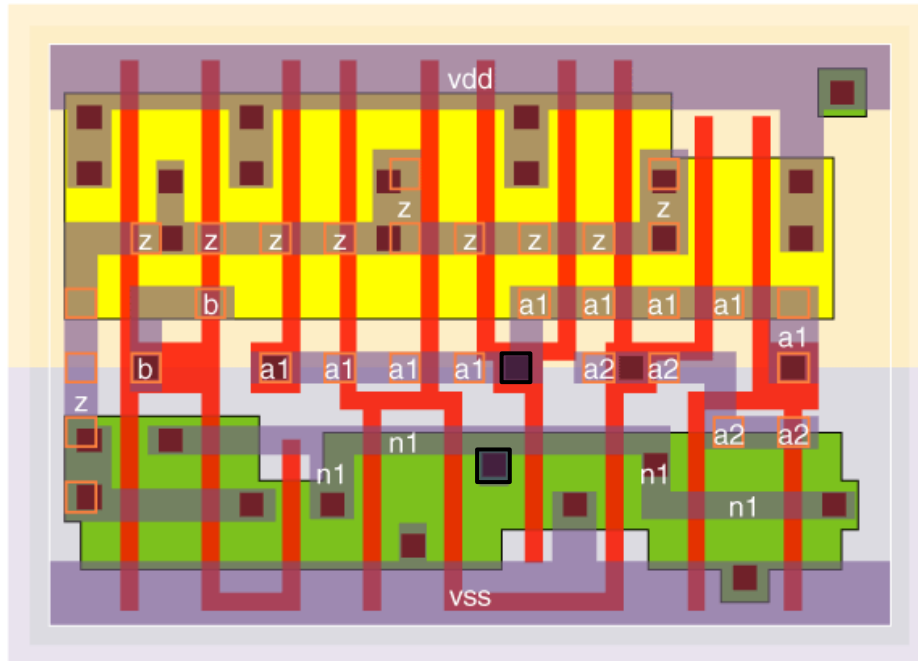
But, connections to power rails are vertical

Canonical Standard Cell M1



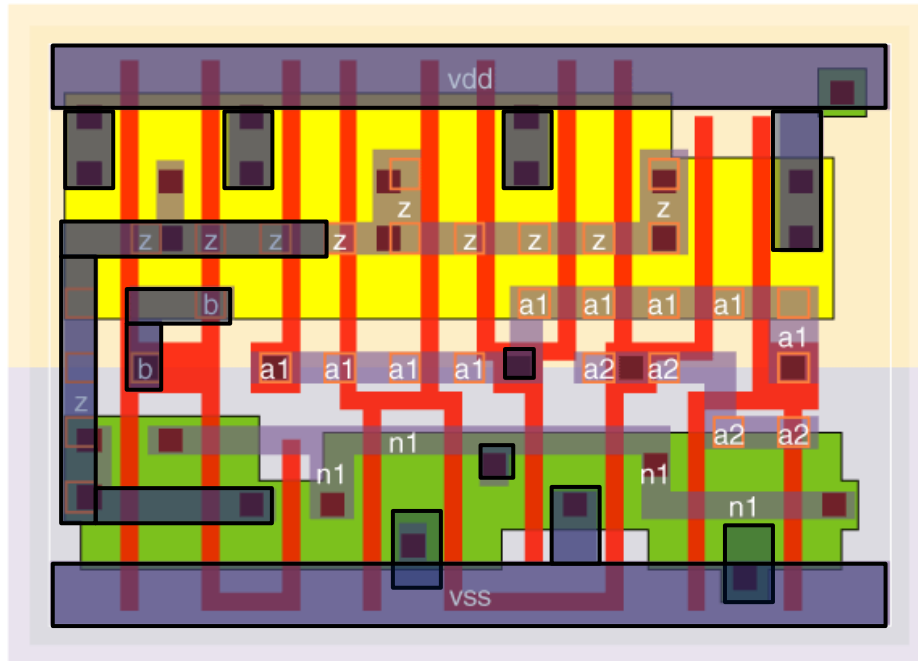
Outputs wrap around inputs
(offset input contact saves space)

Canonical Standard Cell M1



Diffusion and Gate inputs are naturally offset (jogs inherent)

Canonical Standard Cell M1

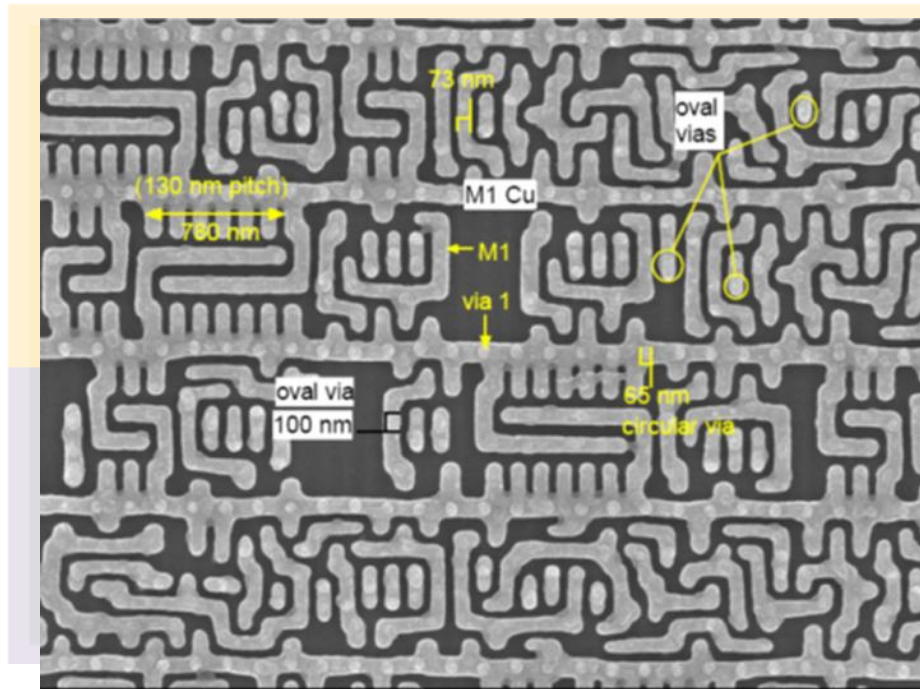


**Requires complex
2-dimensional layout!**

**Jogs
Corners
Tips and Sides**

Canonical Standard Cell M1

as seen in practice:

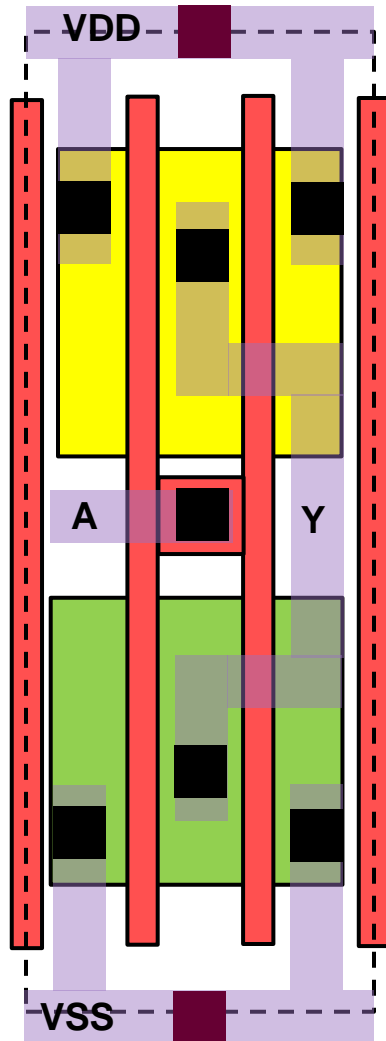


<http://www.chipworks.com/blog/technologyblog/2012/07/31/samsung-32-nm-technology-looking-at-the-layout/>

**Requires complex
2-dimensional layout!**

**Jogs
Corners
Tips and Sides**

Standard Cell Lingo

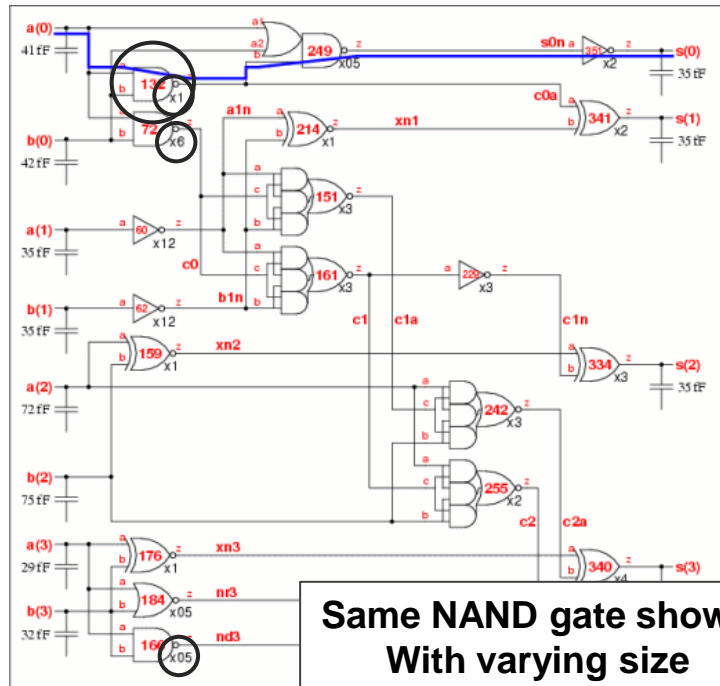
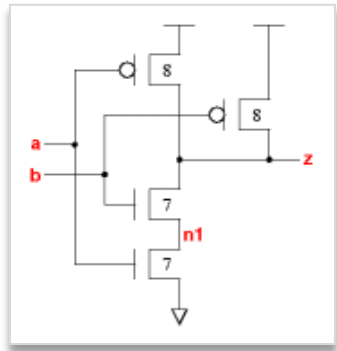


“folded” transistor

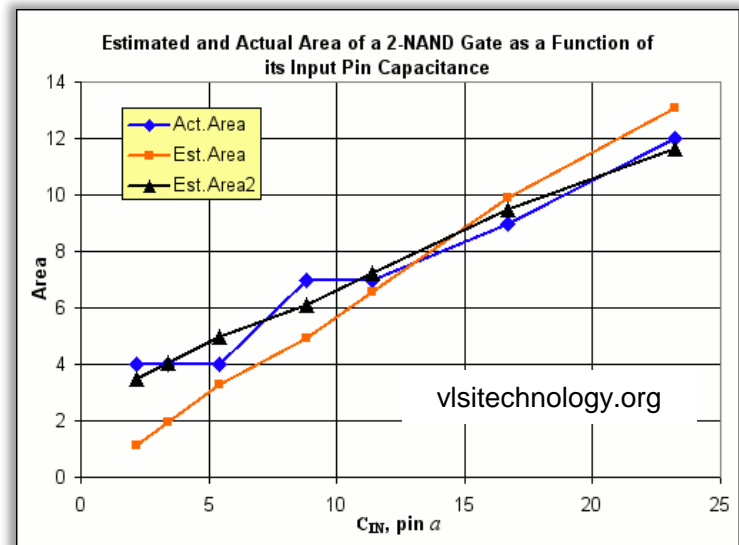
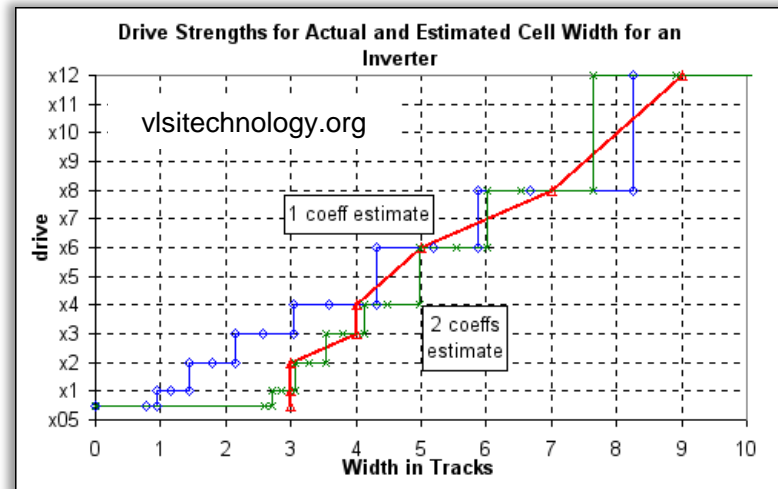
Here: 2 gate fingers per FET

This cell is a “2x” strength Inverter

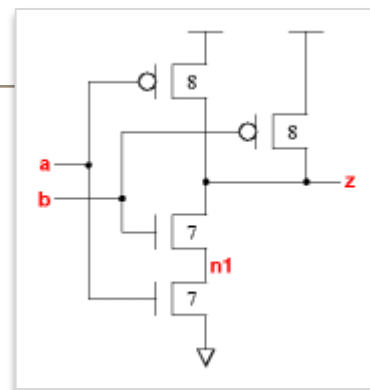
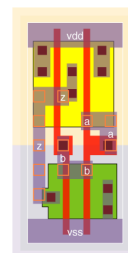
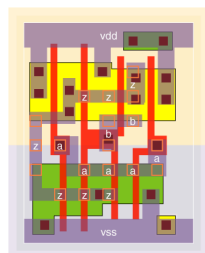
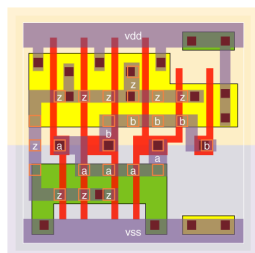
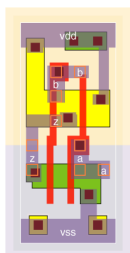
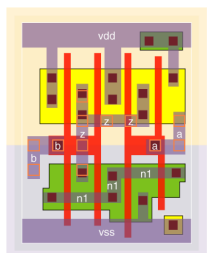
Gate Sizing



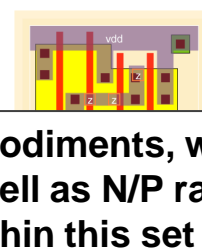
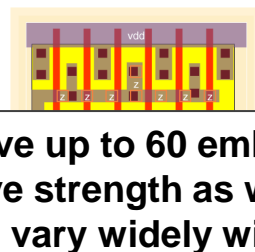
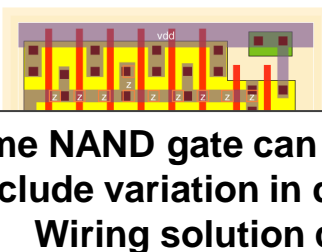
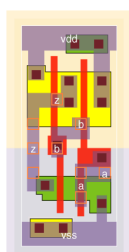
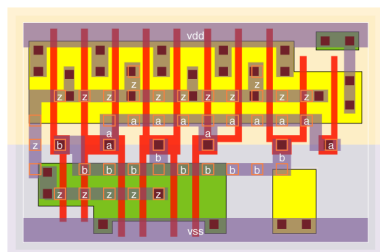
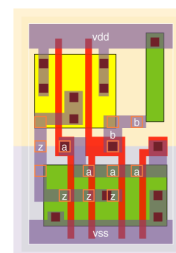
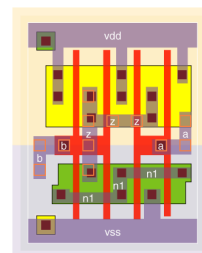
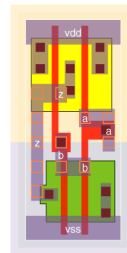
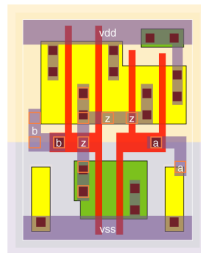
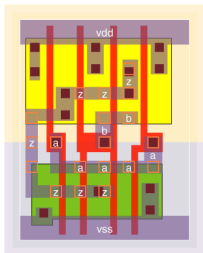
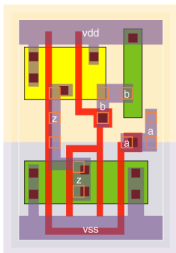
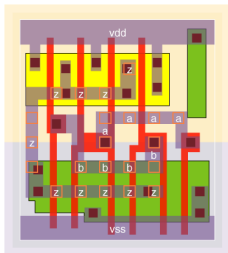
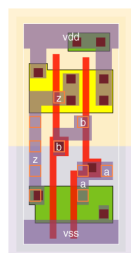
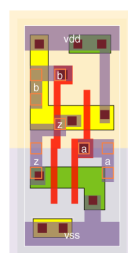
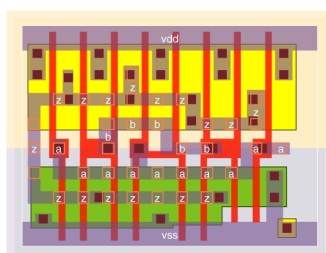
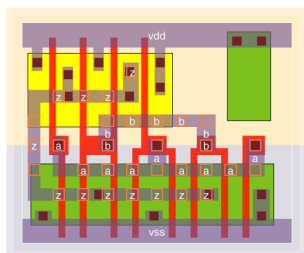
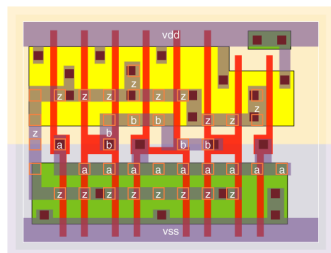
Same NAND gate shown
With varying size



NAND2



vlstechnology.org



Same NAND gate can have up to 60 embodiments, which include variation in drive strength as well as N/P ratio. Wiring solution can vary widely within this set

Standard Cell Library

■ Digital Logic Standard Cells

■ Booleans:

- INV
- NAND
- NOR
- AOI / OAI

Typically 60-70%
of critical paths

■ Complex

- Flip-Flops
- Multiplexers
- Transmission Gates

Typically 30-50%
of logic area

■ Clock, etc.

- Clock/Gated Clock
- Buffers
- Delay
- Level Shifters

Permutations of:

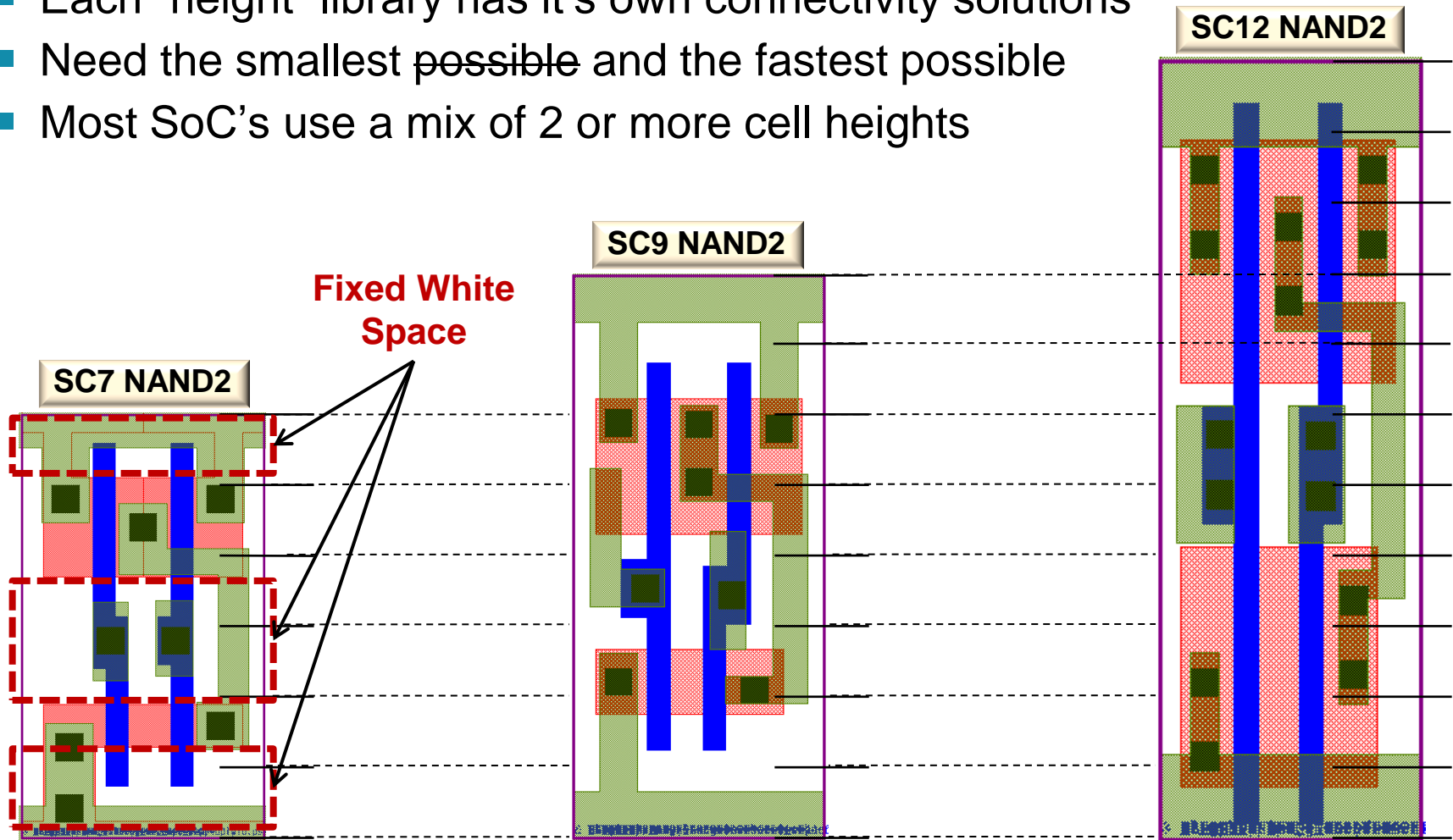
- Beta Ratio
- Drive Strength

**1300-1500 cells
is common**

(per cell height variant)

What Size is Best?

- Each “height” library has it’s own connectivity solutions
- Need the smallest possible and the fastest possible
- Most SoC’s use a mix of 2 or more cell heights



Standard Cells: Summary

- Taking into account the various cell types, drive strengths /beta ratios, and cell heights, there are tens of thousands of unique topologies possible
- All of them must abut arbitrarily to one another, mirrored and offset
- Jogs, tip-to-sides, tip-to-tips, and generally all other kinds of 2D layout are utilized to optimize the solution
- Pin accessibility matters

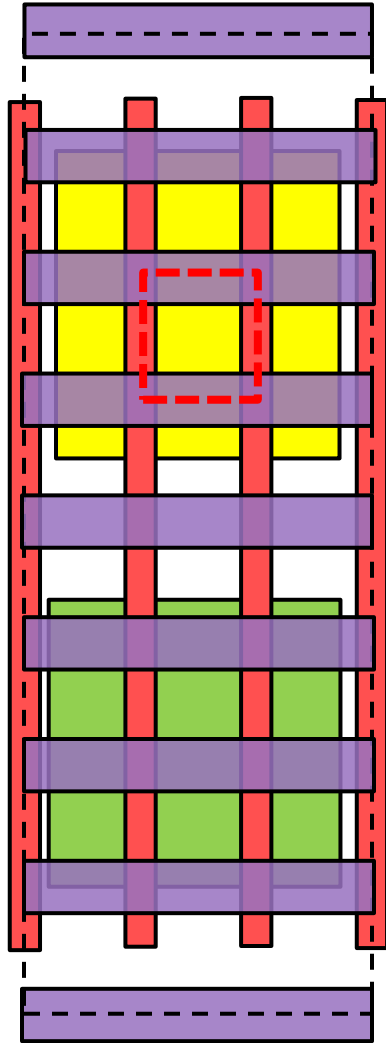
Outline

- | | |
|--|--------|
| 1. Introduction and scope | 5 min |
| 2. SoC Design Overview | 10 min |
| (you are here) | |
| 3. Lithography-Design Interactions and Cost | 25 min |
| <ul style="list-style-type: none">■ History lesson: Do not forget the past■ The messiness of Design-Litho interactions and cost | |

The “do-not-ignore-the-past” section



Standard Cell Size

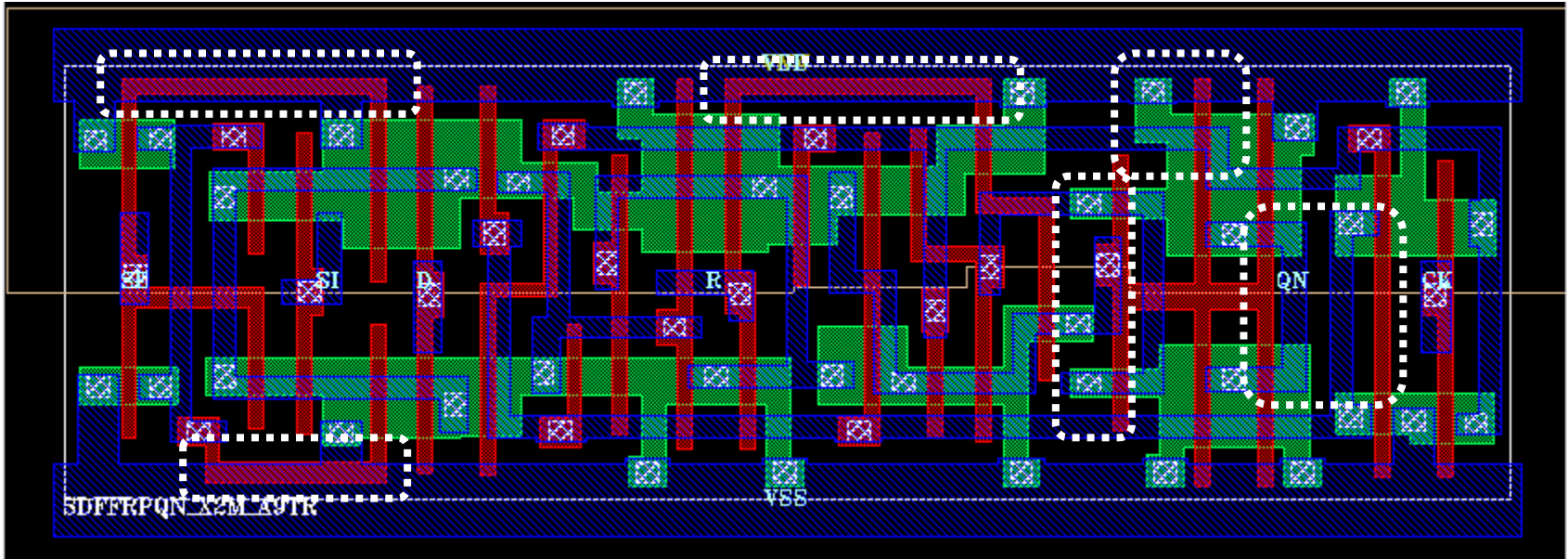


The logic SHOULD scale as:
 $\text{CPP} \times \text{Metal Pitch}$

Technologies are still advertised with
these two numbers

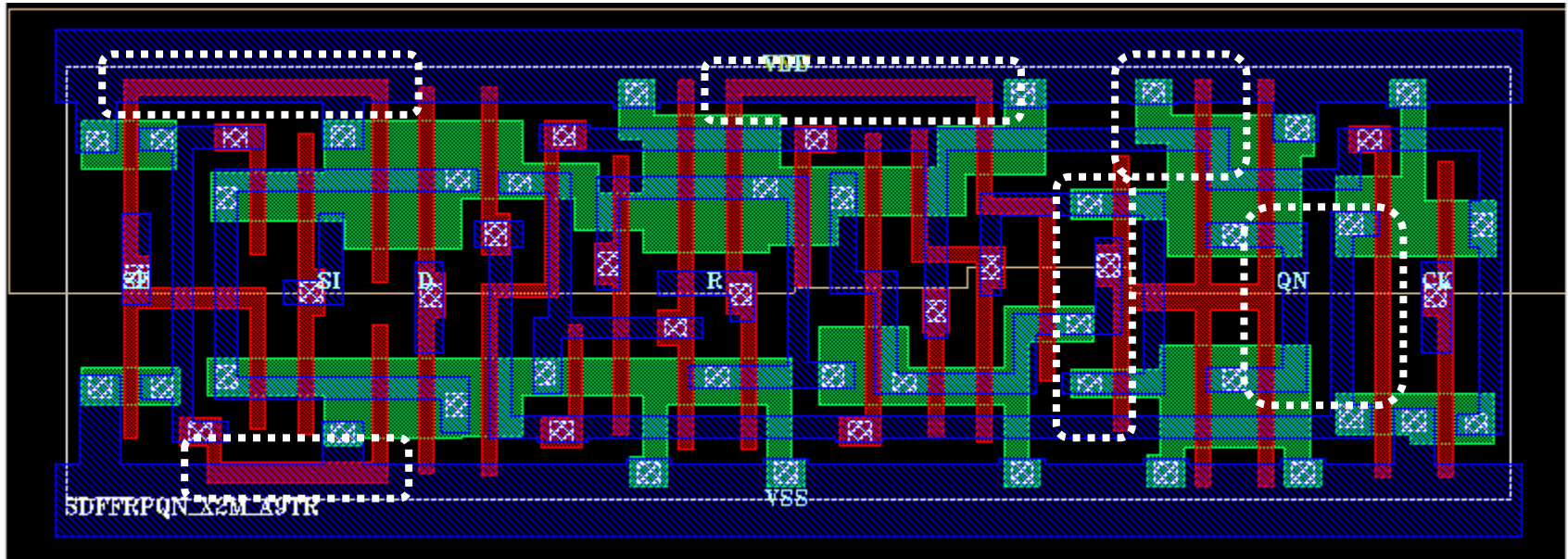
However, logic scaling is far more
complicated (and almost always less than
the pitch scaling entitlement)

65nm flip flop



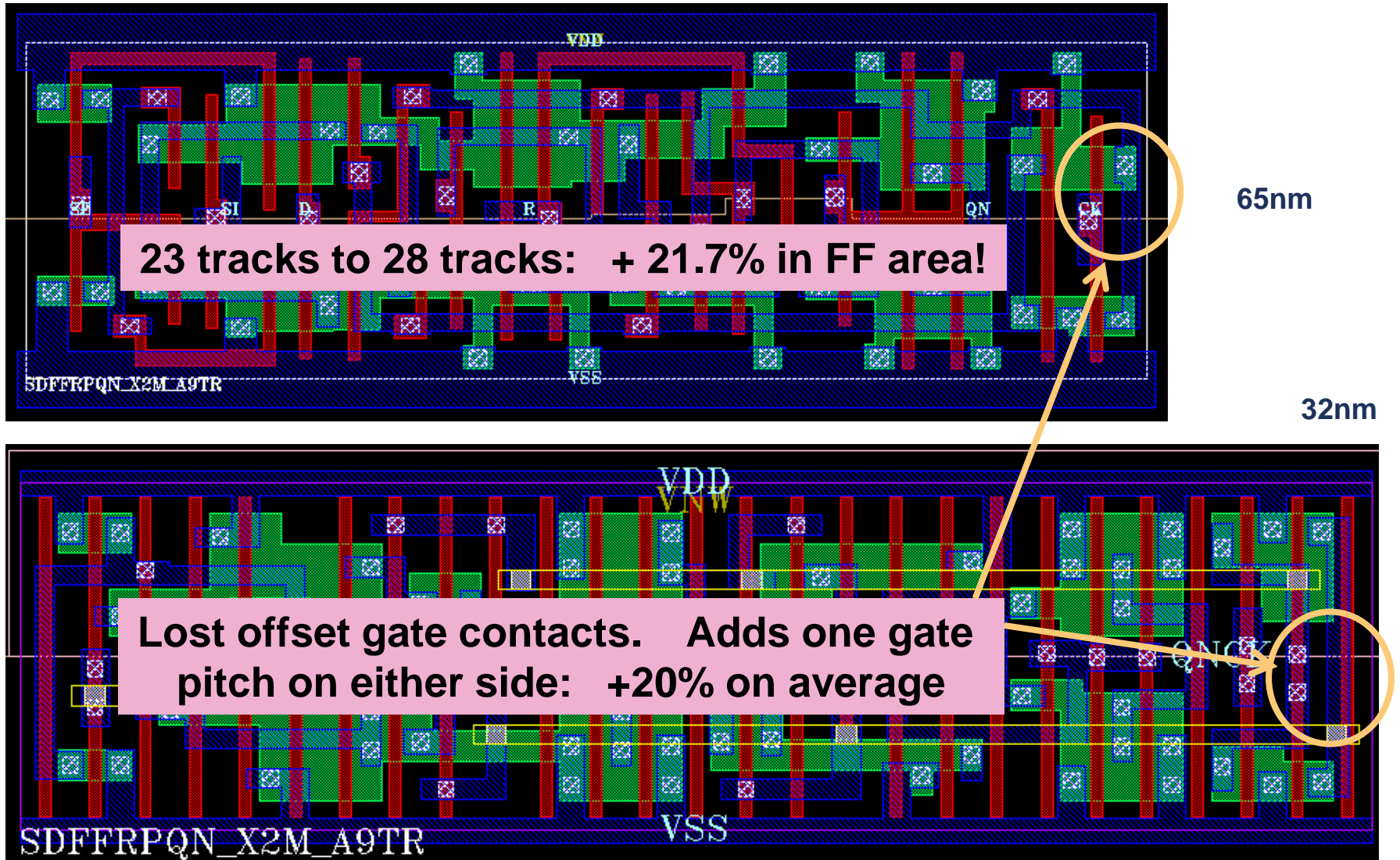
- Why DFM was invented
- Great gate density, all of the jumbled layout has a purpose:
 - Diffusion tabs get power to the FETs without blocking M1
 - Contact pitch not limiting anything: 3 independent RX contacts in one PC pitch
 - Horizontal PC routes, including “outbound” routes under the M1 rails
 - Uneven gate pitch allows some creative routing
 - M1 tips/sides everywhere

Logic area $\propto 1/k_1$

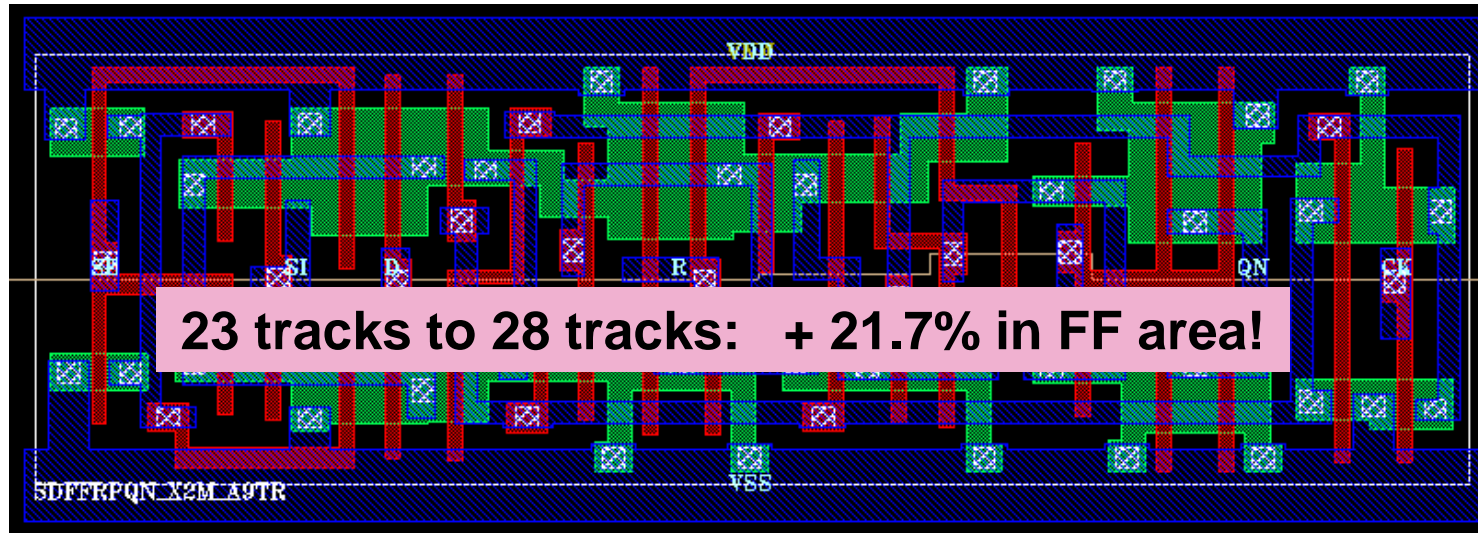


- But none of these tricks are legal anymore
 - Diffusion tabs lost
 - Gate tabs and bends became too expensive due to rounding (via the corner models)
 - Gate pitch uniformity also enforced via proximity rules (via the corner models)

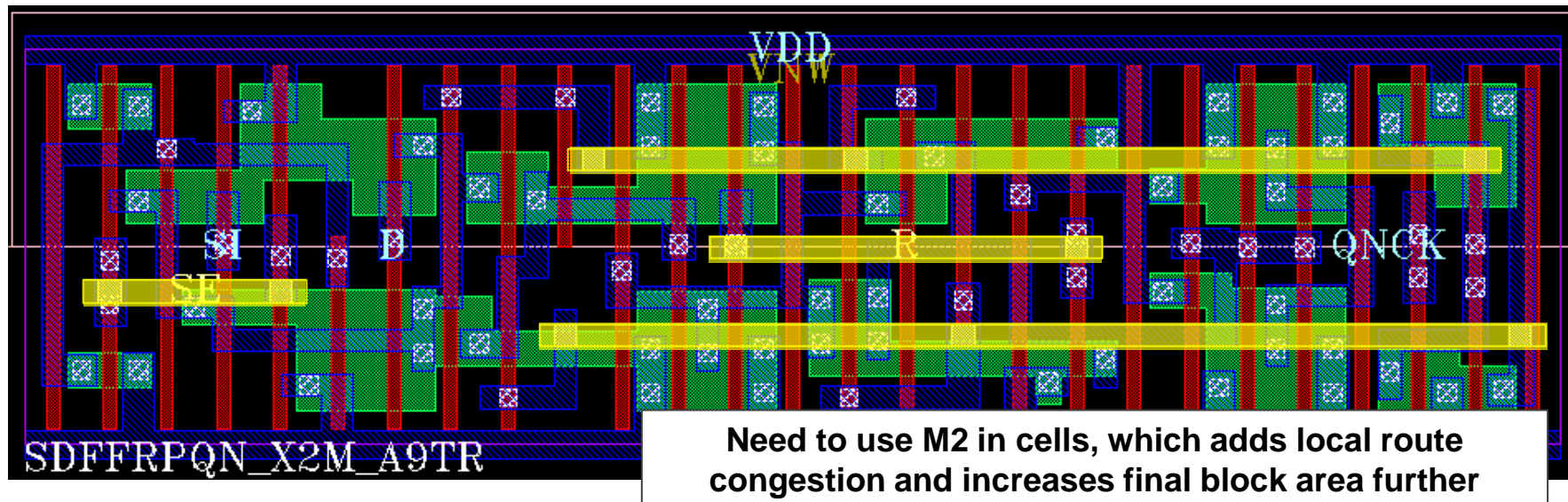
Logic area $\propto 1/k_1$



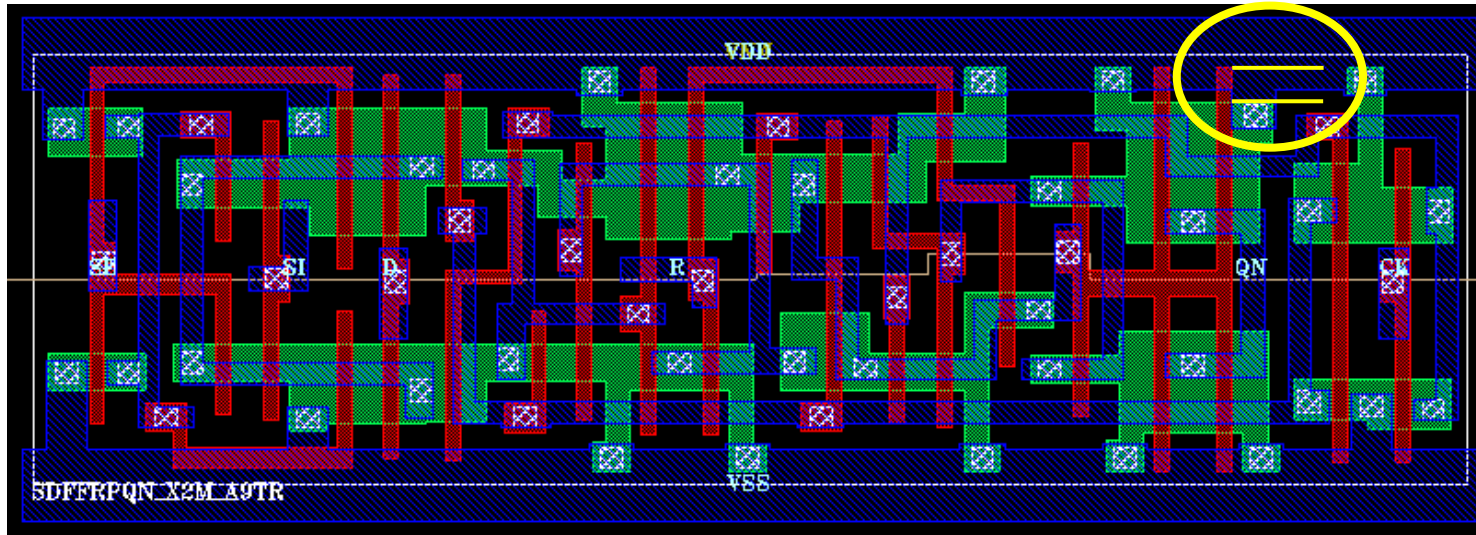
More cost “above the cells”



32nm

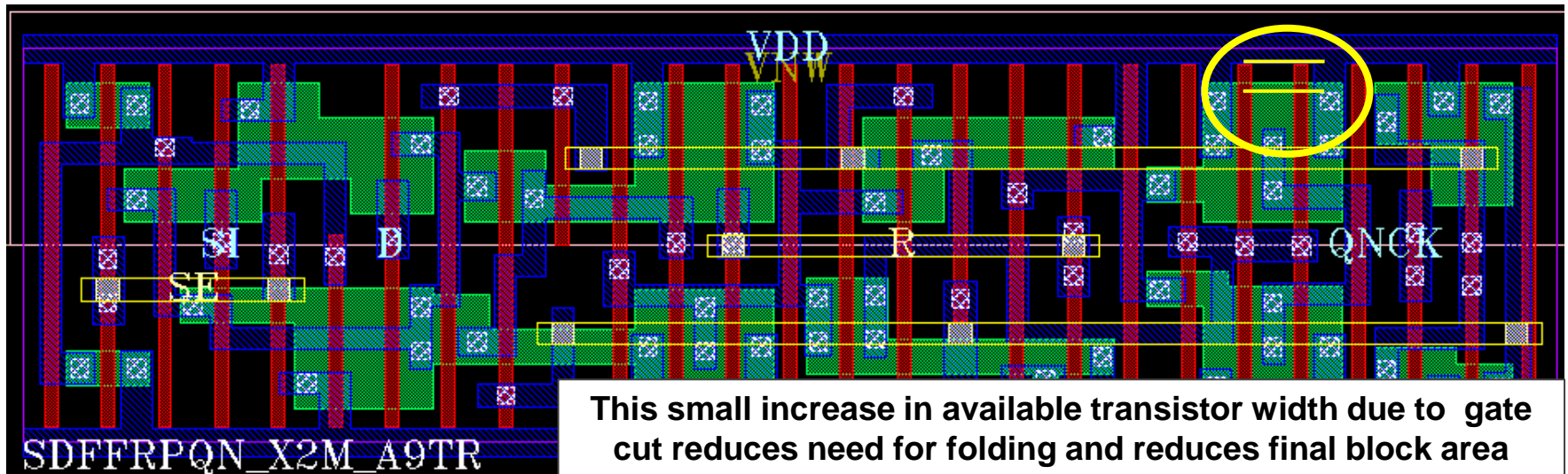


28nm Gate Double patterning (cut)



65nm

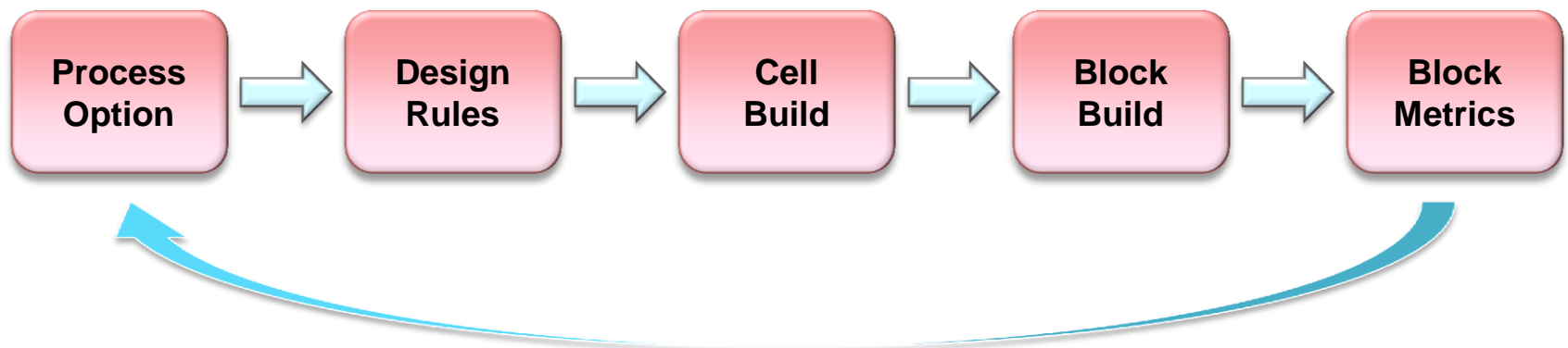
32nm



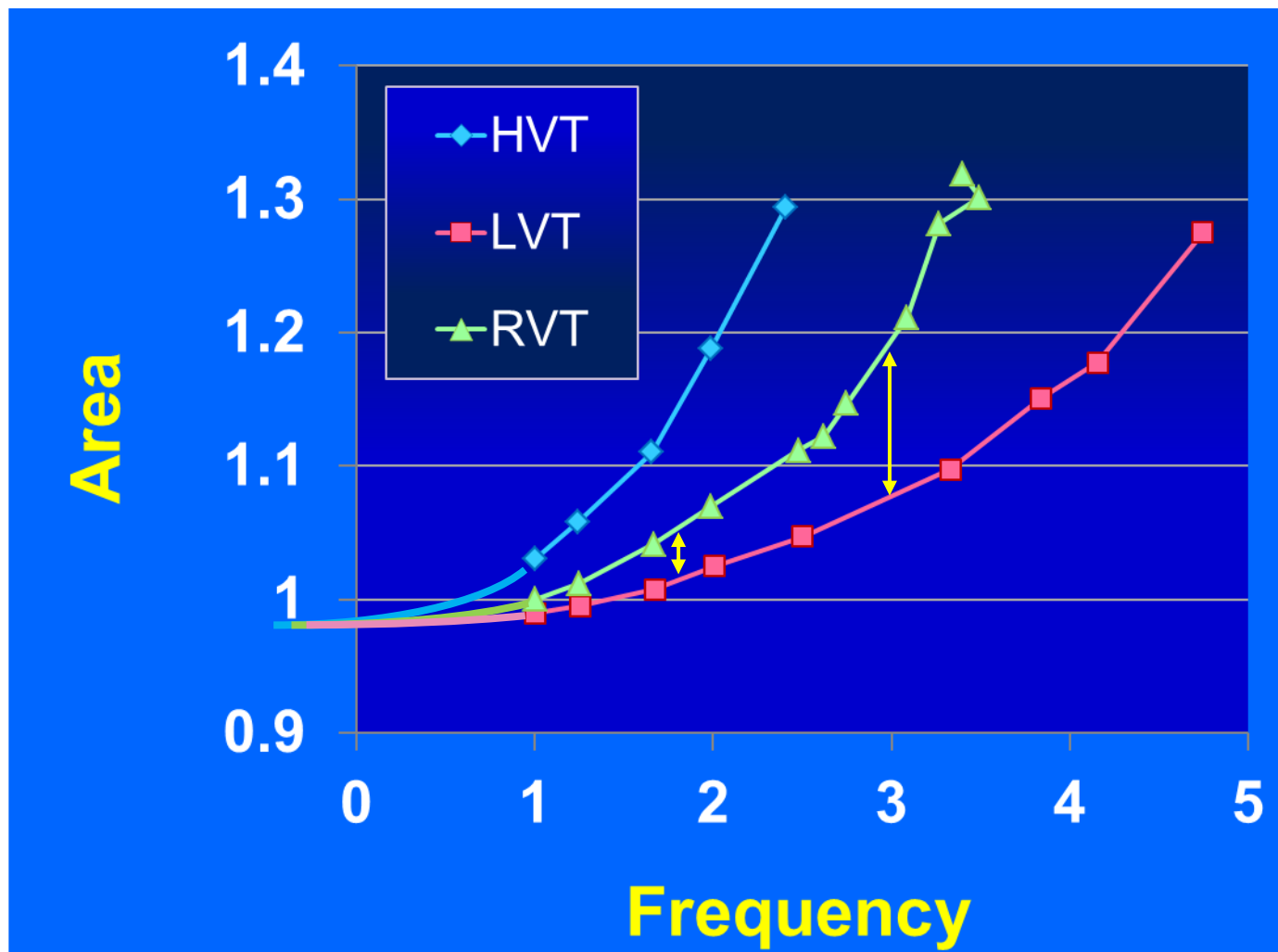
This small increase in available transistor width due to gate cut reduces need for folding and reduces final block area

Why it is hard to evaluate specific rules

- Expensive and time consuming
 - Physical IP creation is very detailed and laborious
 - P&R tools are never ready for generation N+1 lithography rules
- Some of the rectangles define transistors
 - To evaluate the system, we also need device models
- Drive strength must be evaluated in the context of wiring parasitics



Block Area vs. FET Strength



Same circuit ends up physically larger if you end up with weaker transistors.
Actual amount of difference depends on the frequency target specified for SP&R

Public FinFET Models

Predictive
Technology
Model

Introduction

Latest Models

Nano-CMOS

Post-Silicon

Interconnect

Reliability

Contact



LATEST MODELS

<http://ptm.asu.edu/>

Typical SPICE model files for each future generation are available here.

Attention: By using a PTM file, you agree to acknowledge both the URL of PTM: <http://ptm.asu.edu/> and the related public

New!

June 01, 2012:

PTM releases a new set of models for multi-gate transistors (PTM-MG), for both HP and LSTP applications. It is based on BSIM

Acknowledgement: PTM-MG is developed in collaboration with ARM.

Please start from [models](#) and [param.inc](#).

- 7nm PTM-MG [HP NMOS](#), [HP PMOS](#), [LSTP NMOS](#), [LSTP PMOS](#)
- 10nm PTM-MG [HP NMOS](#), [HP PMOS](#), [LSTP NMOS](#), [LSTP PMOS](#)
- 14nm PTM-MG [HP NMOS](#), [HP PMOS](#), [LSTP NMOS](#), [LSTP PMOS](#)
- 16nm PTM-MG [HP NMOS](#), [HP PMOS](#), [LSTP NMOS](#), [LSTP PMOS](#)
- 20nm PTM-MG [HP NMOS](#), [HP PMOS](#), [LSTP NMOS](#), [LSTP PMOS](#)

The entire package is also available here: [PTM-MG](#)

November 15, 2008:

PTM releases a new set of models for low-power applications (PTM LP), incorporating high-k/metal gate and stress effect.

- 16nm PTM LP model: [V2.1](#)
- 22nm PTM LP model: [V2.1](#)
- 32nm PTM LP model: [V2.1](#)
- 45nm PTM LP model: [V2.1](#)

September 30, 2008:

PTM releases a new set of models for high-performance applications (PTM HP), incorporating high-k/metal gate and stress effect.

- 16nm PTM HP model: [V2.1](#)
- 22nm PTM HP model: [V2.1](#)

Some free FinFET models you can use for evaluations
such as prior slide

CMC-endorsed BSIM-MG (HSPICE LEVEL=72, e.g.)

Exploring Sub-20nm FinFET Design with Predictive Technology Models

Saurabh Sinha, Greg Yeric, Vikas Chandra, Brian Cline, Yu Cao*
ARM Inc., *Arizona State University, Tempe, AZ
saurabh.sinha@arm.com

ABSTRACT

Predictive MOSFET models are critical for early stage design-technology co-optimization and circuit design research. In this work, Predictive Technology Model files for sub-20nm multi-gate transistors have been developed (PTM-MG). Based on MOSFET scaling theory, the 2011 ITRS roadmap and early stage silicon data from published results, PTM for FinFET devices are generated for 5 technology nodes corresponding to the years 2012-2020 on the ITRS roadmap.

Categories and Subject Descriptors

B.7.1 [Hardware]: Integrated Circuits—Types and Design Styles, Advanced Technologies; B.8.2 [Hardware]: Performance and Reliability—Performance Analysis and Design Aids

General Terms

Theory

Keywords

FinFET, multi-gate, scaling theory, predictive models, SPICE

1. INTRODUCTION

CMOS scaling has continued up to the 20nm node through innovative techniques such as incorporating high-k dielectrics in the gate stack, strain engineering, pocket implants and optimization in materials and device structures. However, further scaling of planar devices is proving to be extremely challenging due to degrading short channel effects, process variations and reliability degradation [1].

Multi-gate transistor structures such as FinFETs will be the technology of choice for extending CMOS scaling beyond the 20nm node. Improved short channel control through a fully depleted fin, reduced random dopant fluctuation, improved mobility, lower parasitic junction capacitance and improved area efficiency are some of the primary advantages

of FinFETs [2]. However, FinFETs will be markedly different than planar FETs due to added fringing capacitance, higher access resistance, width-quantization, 3D-factor, and

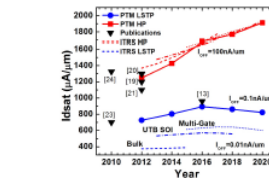


Figure 8: Prediction of I_{DSAT} for PTM-MG HP and LSTP devices. Trends from ITRS and selected publications are annotated for comparison.

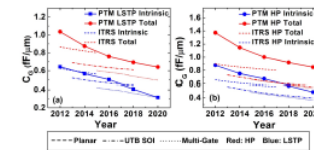


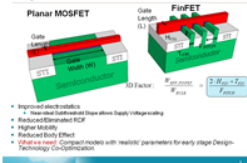
Figure 10: Intrinsic and total gate capacitance of PTM-MG LSTP and HP FinFETs. ITRS prediction trends are plotted for comparison.

2012 and 2014 PTM-MG devices are made to approximate

Exploring Sub-20nm FinFET Design with Predictive Technology Models

Saurabh Sinha, Greg Yeric, Vikas Chandra, Brian Cline, Yu Cao*
ARM Inc., *Arizona State University, Tempe, AZ

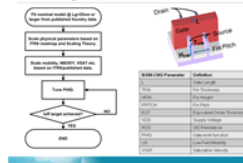
Why FinFETs?



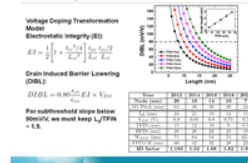
BSIM-CMG for FinFETs

- Compact model for common multi-gate FETs developed by BSIM Group at UC Berkeley.
- Device device data (https://www.bsim.com/bsim4/bsim4cmg/)
- Compatible with commercial circuit simulators such as HSPICE (Level-72).
- Endorsed by Compact Modeling Council (CMC) as industry standard model for FinFETs.
- Captures real-device effects such as:
 - Geometry based parasitic fringing capacitance
 - Geometry based parasitic access resistance
- PTM-MG Predictive Technology Models for Multi-gate transistors consists of model parameter cards based on BSIM-CMG.

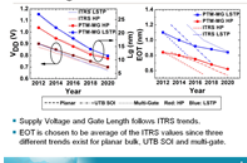
Predicting Model Parameters



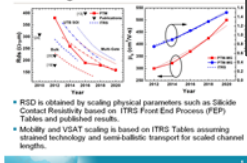
Predicting Model Parameters



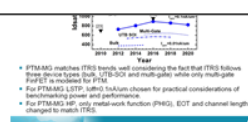
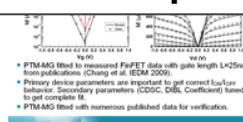
Predicting Model Parameters



Predicting Model Parameters

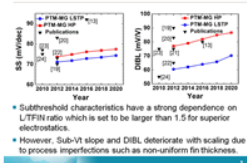


Description of the free FinFET models

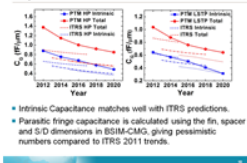


Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
DAC 2012, June 3-7, 2012, San Francisco, California, USA
Copyright 2012 ACM 978-1-4503-1199-1/12/06...\$10.00.

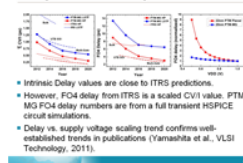
Subthreshold Characteristics



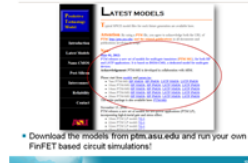
Capacitance



Design Benchmarking



What's next?

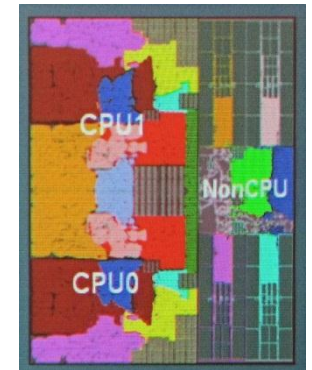
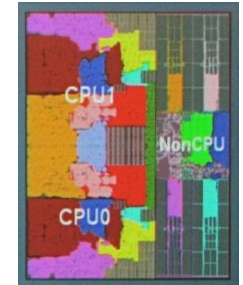


Logic Scaling: 65nm to 28nm

- | | |
|--|---------|
| ■ Loss of active and gate routing | 2 - 5% |
| ■ Loss of gate contact offset | 4 - 8% |
| ■ Loss of M1 auto routing | 5 - 10% |
| ■ Contact and via restrictions | 2 - 5%* |
| ■ M1 tip/side restrictions | 2 - 5%* |
| ■ Increased use of M2 in complex cells | 2 - 5% |
| ■ Poly cut double patterning | - 3% |

*combined, these also include minimum cell height scaling penalty of 9/8

Actual results are highly dependent on the library quality, implementation quality, design and design targets, so focus on overall issue set and trends.

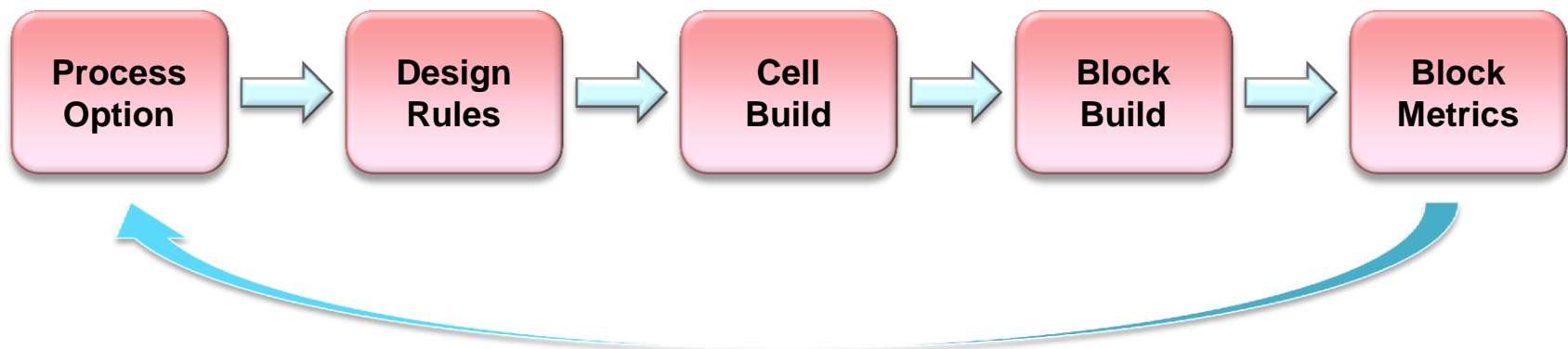


65 to 45 to 28:

Along the way we lost a half node of area scaling in the logic!
("mitigated" by local wire congestion)

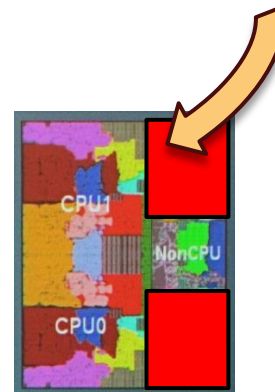
Cost: EUV vs. The Alternatives

- In terms of design interactions:
 - Don't ignore the past!
 - Cost is often viewed as relative to the previous node, but over time small cost adders can add up
- The interaction with design is complex and varied. You must understand the application and flow, and then go through it, in order to see and quantify all of the issues.
 - Things are too complex to just look at one NAND gate
 - The answer always depends on the circuit and the targets

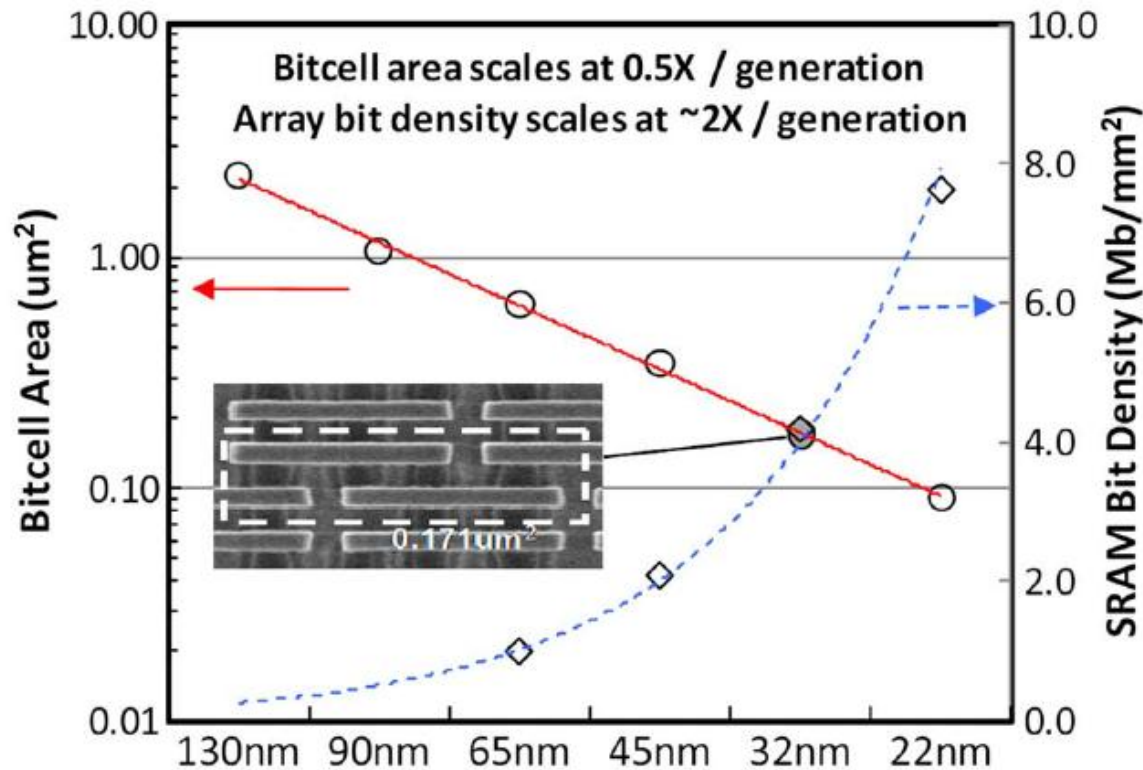


SoC Contents

~ 20% I/O
~ 40% logic
~ **40% memory**

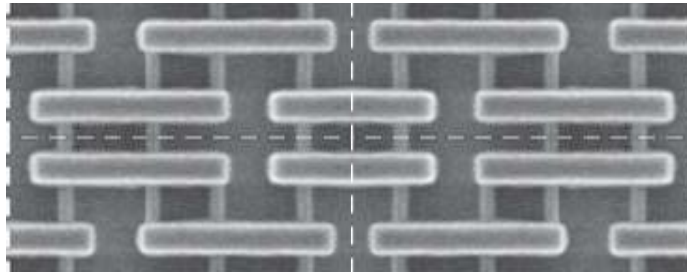


Advertised SRAM Bitcell Area Trend

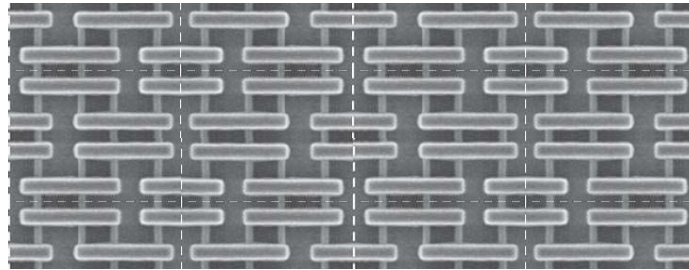


Karl, E., et al., IEEE JSSC, Jan 2011, pp 76-84

2 x 2 bitcell array

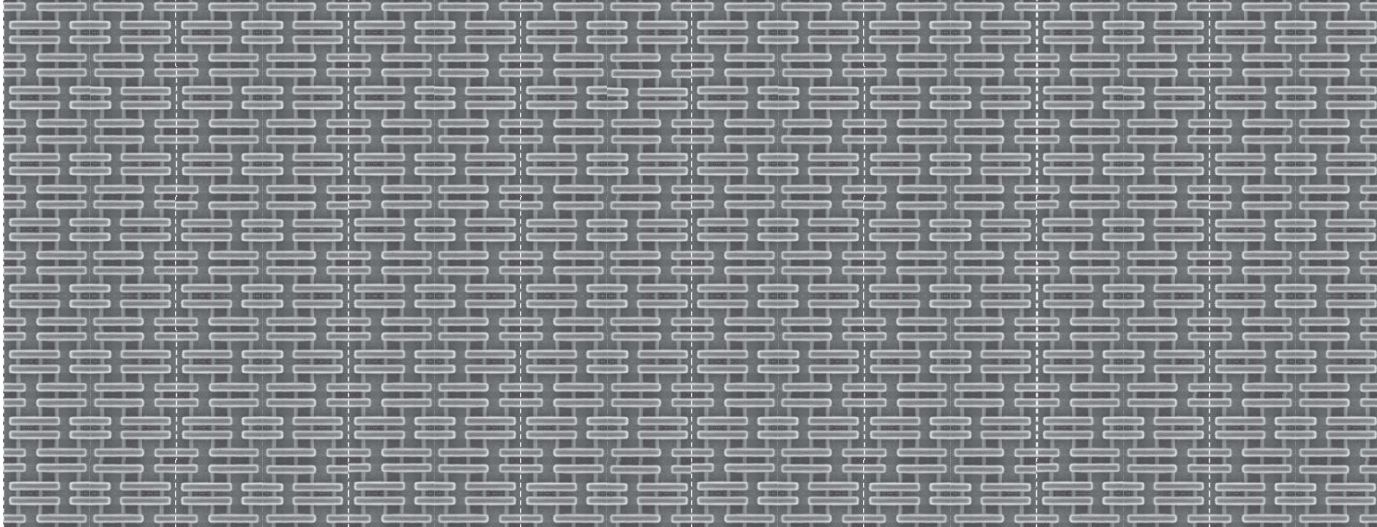


4 x 4 bitcell array



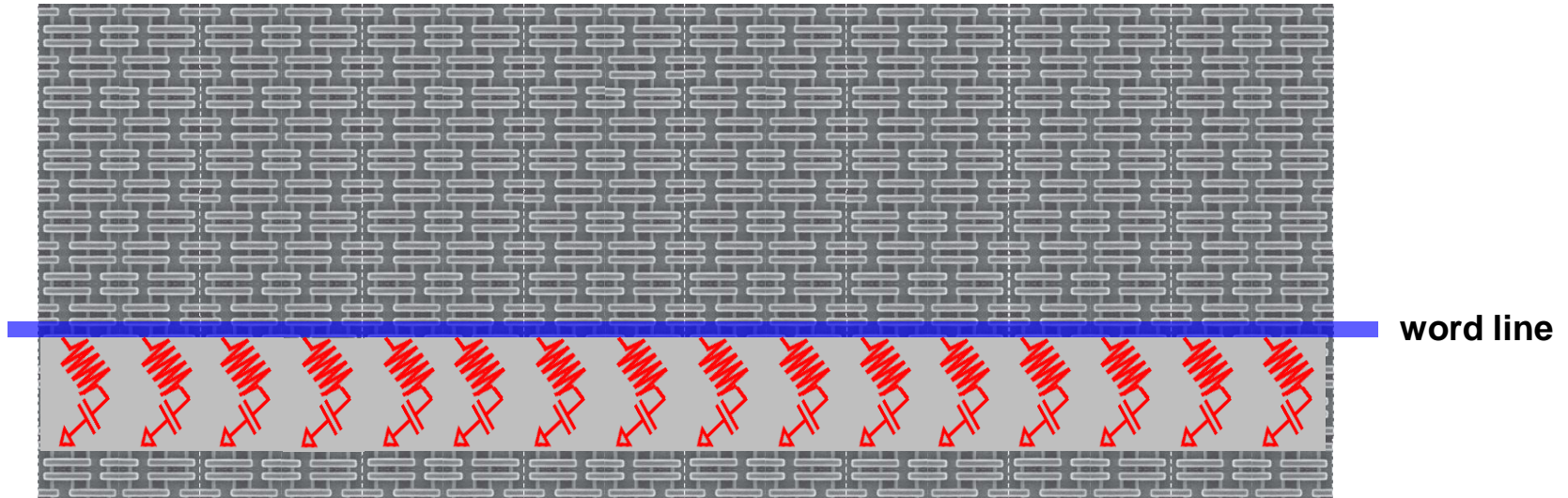
Replicated arrays of unit cells end up looking nice to lithographers

16 x 16 bitcell array



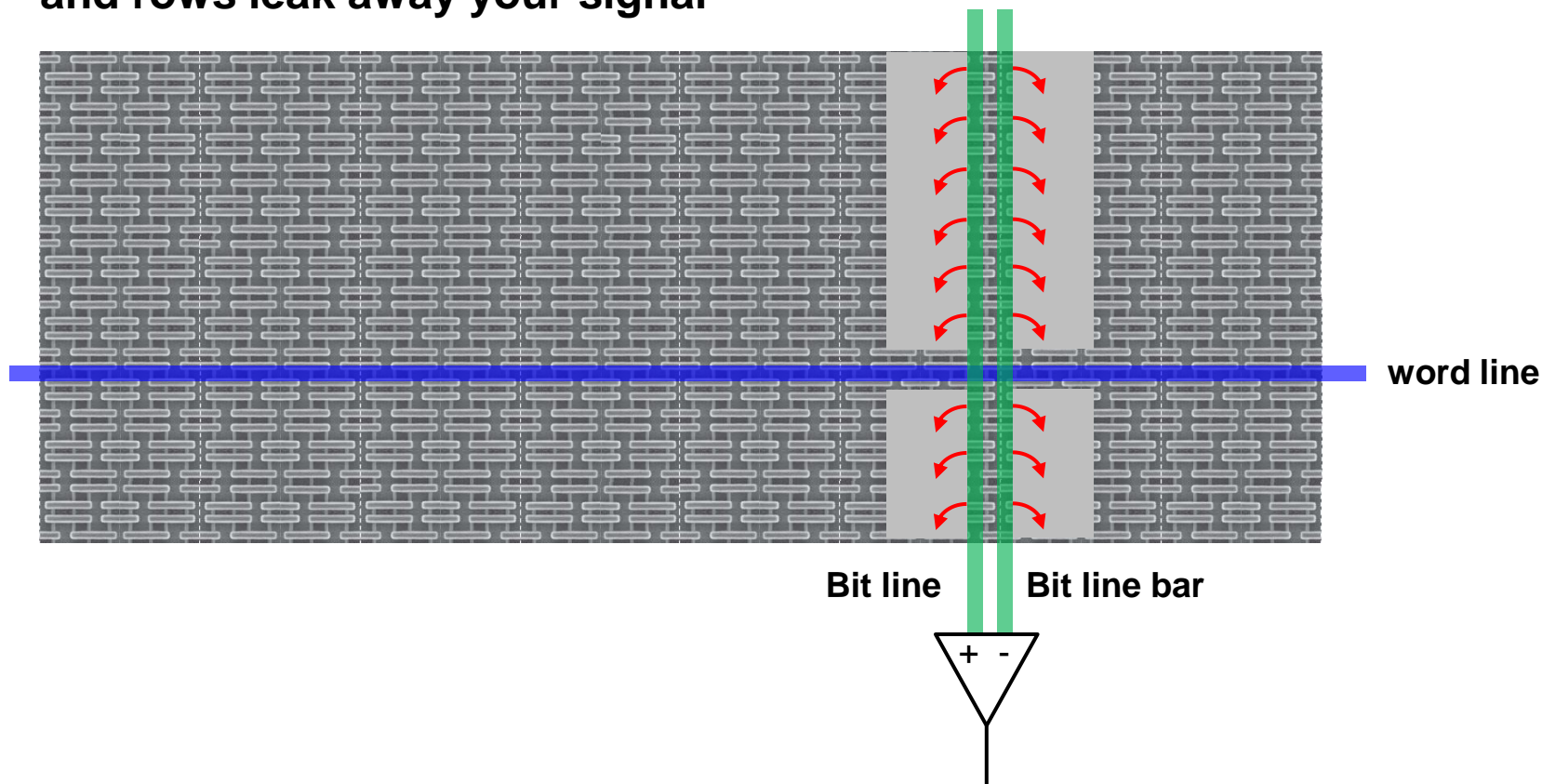
16 x 16 bitcell array

Columns slow you down,



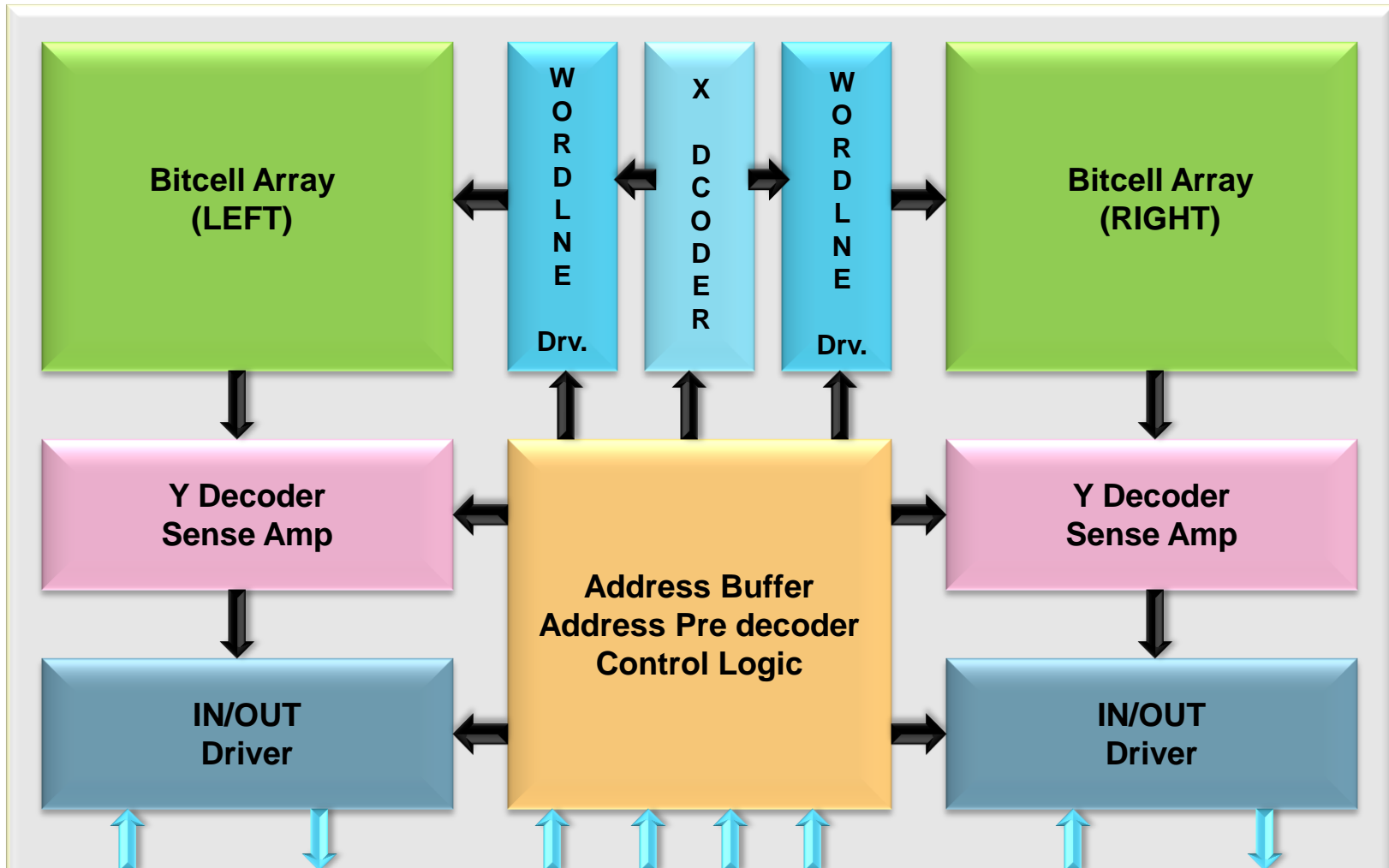
16 x 16 bitcell array

Columns slow you down,
and rows leak away your signal



Maximum “bank” size: 16 to 256 rows/columns

SRAMs Row/Column Limits



Bank size of arrays limited to previous issues, meaning logic overhead is fixed to significant fractions of the overall memory

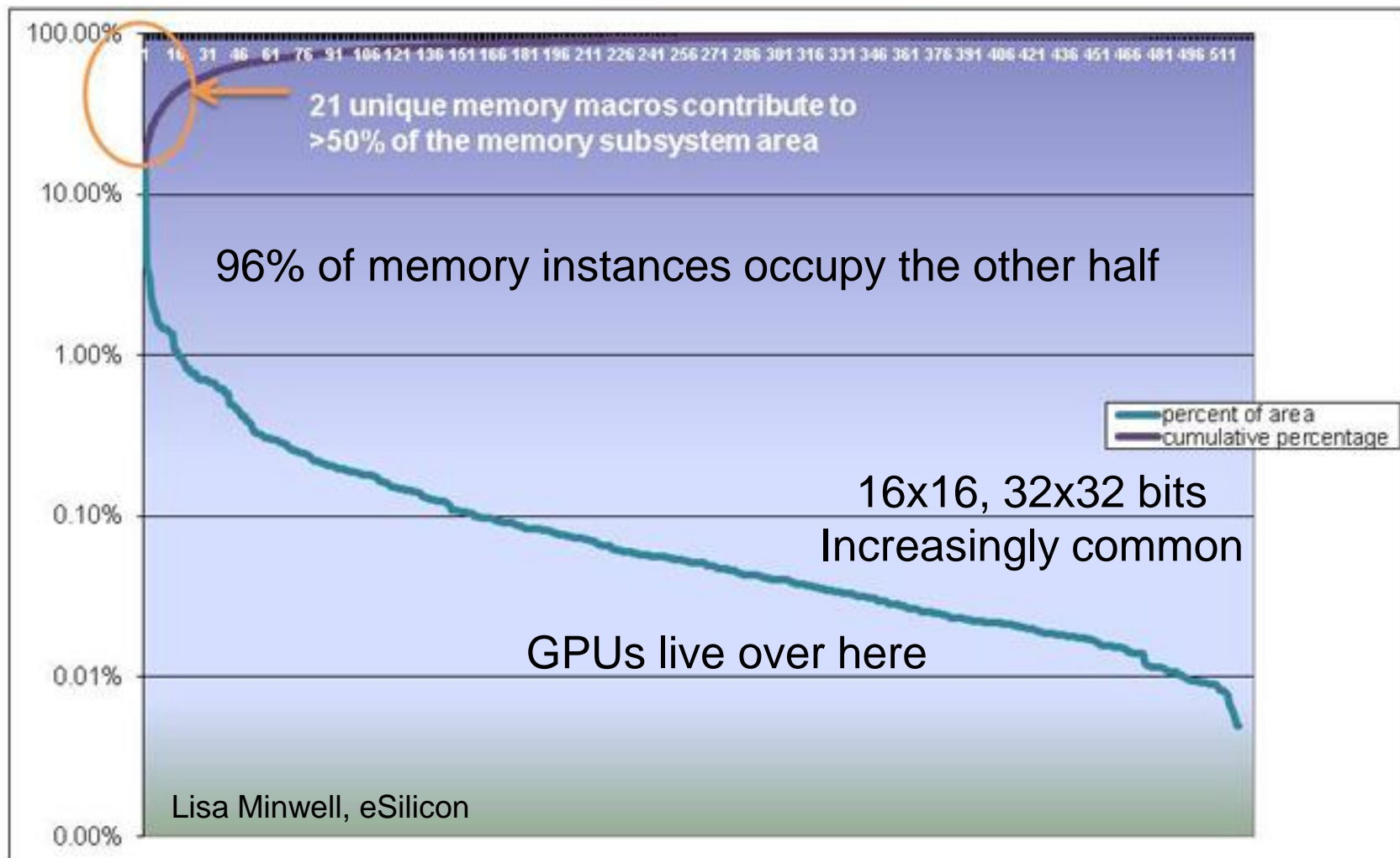
SRAMs Row/Column Limits

This memory instance has a bitcell efficiency of about 50% (and this is a good example)



www.arm.com

Memory Instances in SoC



Moreover, most SoCs are dominated by small memory instances

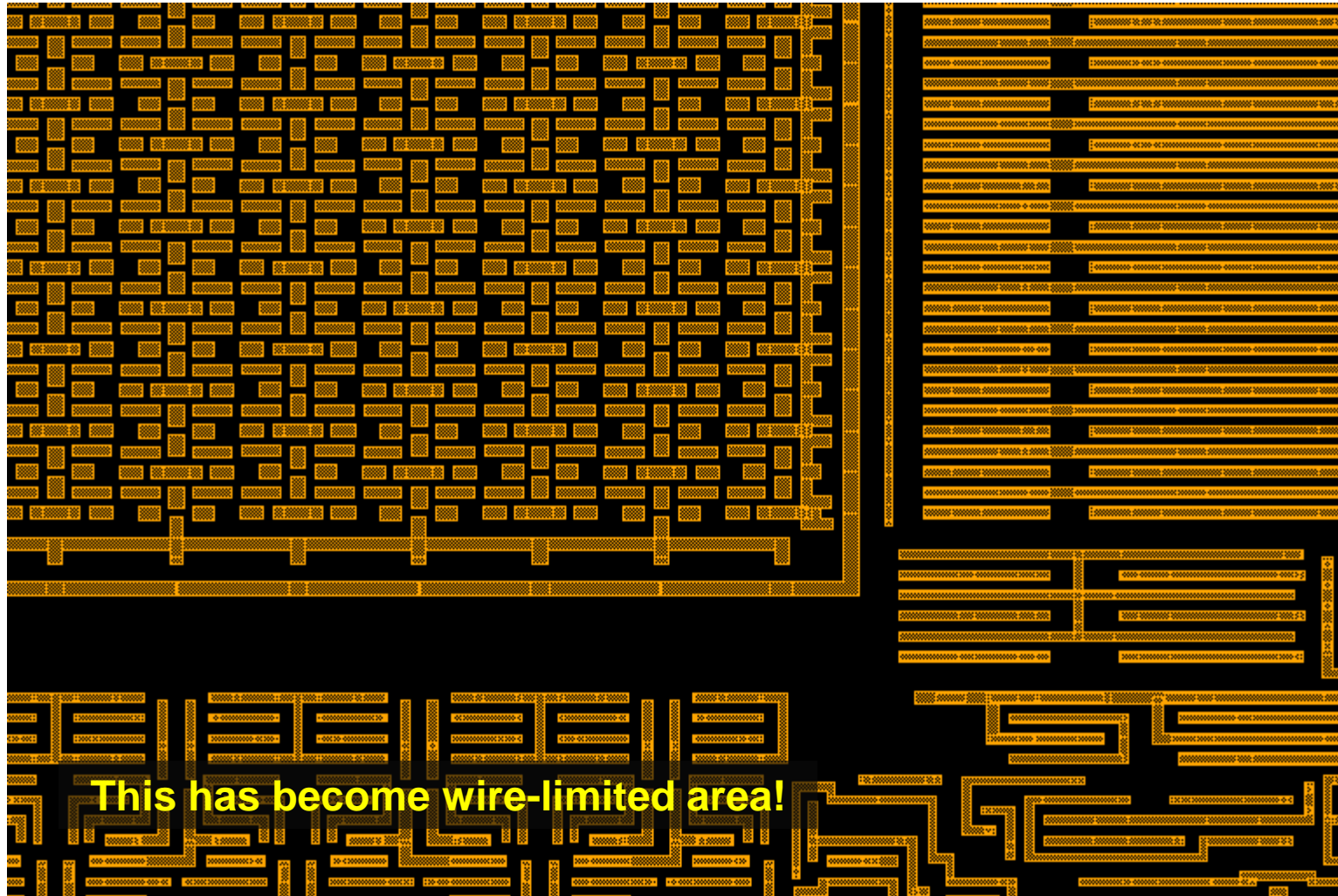
SRAM Array-Periphery Transition

- Owing to bank sizes and small instances, transition region is important to keep small



SRAM Array-Periphery Transition

Metal example:



SRAM Array-Periphery Transition

- Note need for jogs when pitch-matching bitcell array to periphery (some multiple patterning techniques can not support this)



SRAM Array-Periphery Transition

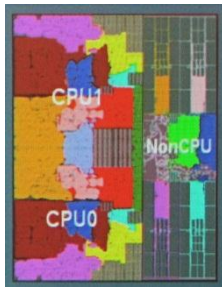


SoC SRAM Composition Summary

- ~ 40% bitcell efficiency
- Rest scales more like logic
- The trend is for smaller bank sizes and more overhead (design assist)
- Bitcell array transition to logic is not litho-friendly (and is different vertically vs. horizontally)
 - Requires increasing cost in transition dead space
 - Dummy end cap cells, e.g.

SoC Scaling: 65nm to 32/28

Lithographically-limited issues:



~ 20% I/O

~ 40% logic

~ 40% memory

~ 30% under-scaled

~ 12% under-scaled

SoC Scaling: 32/28nm to EUV

- Below 22nm, contact (holes) and metal pitches push to below their single-patterning limits (using foundry metal stack)
- With today's technology, that means multiple patterning steps
 - 20 nm: 64 nm M1 pitch, DP OK
 - 14 nm (scaled from 20): 48 nm M1 pitch, need TP/QP
 - Contacts need to be able to make the gate pitch
 - Vias need to hit the diagonal pitch of their respective layers
 - 80nm pitch diagonal: 113 nm
 - 64nm pitch diagonal: 96 nm

Canonical 32nm gate contacts

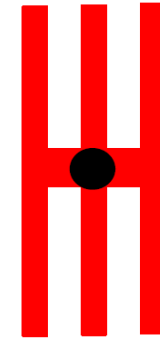
3 successive
Independent
gate contacts



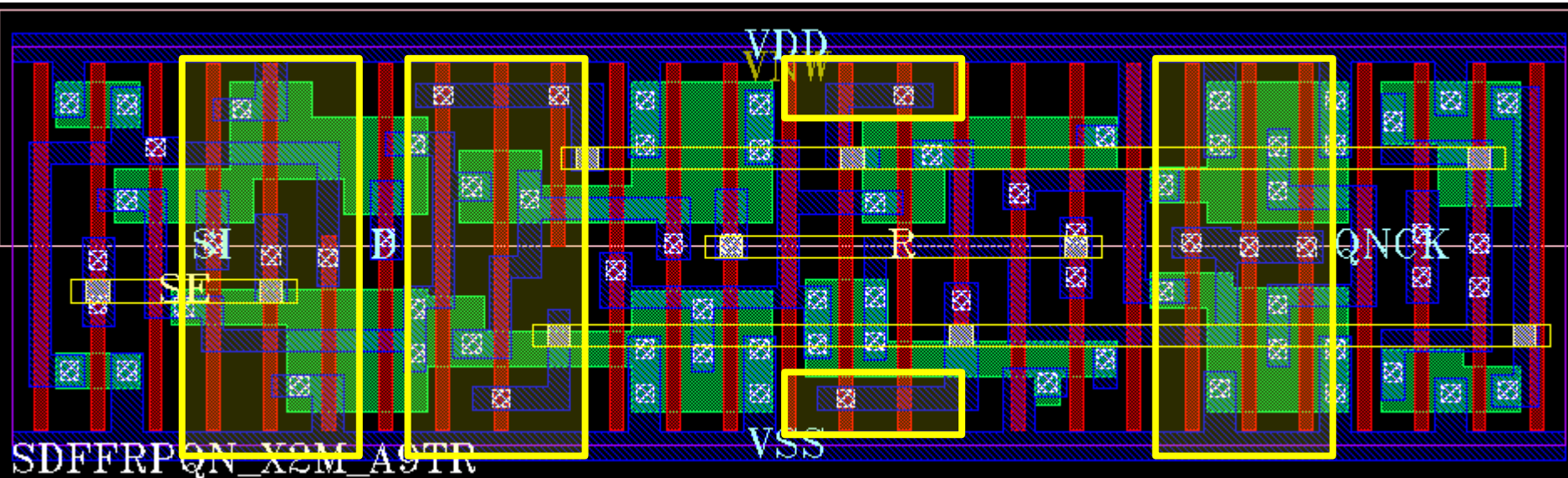
Cross-coupled
Gate contact



Gate contacts
from outer
channels

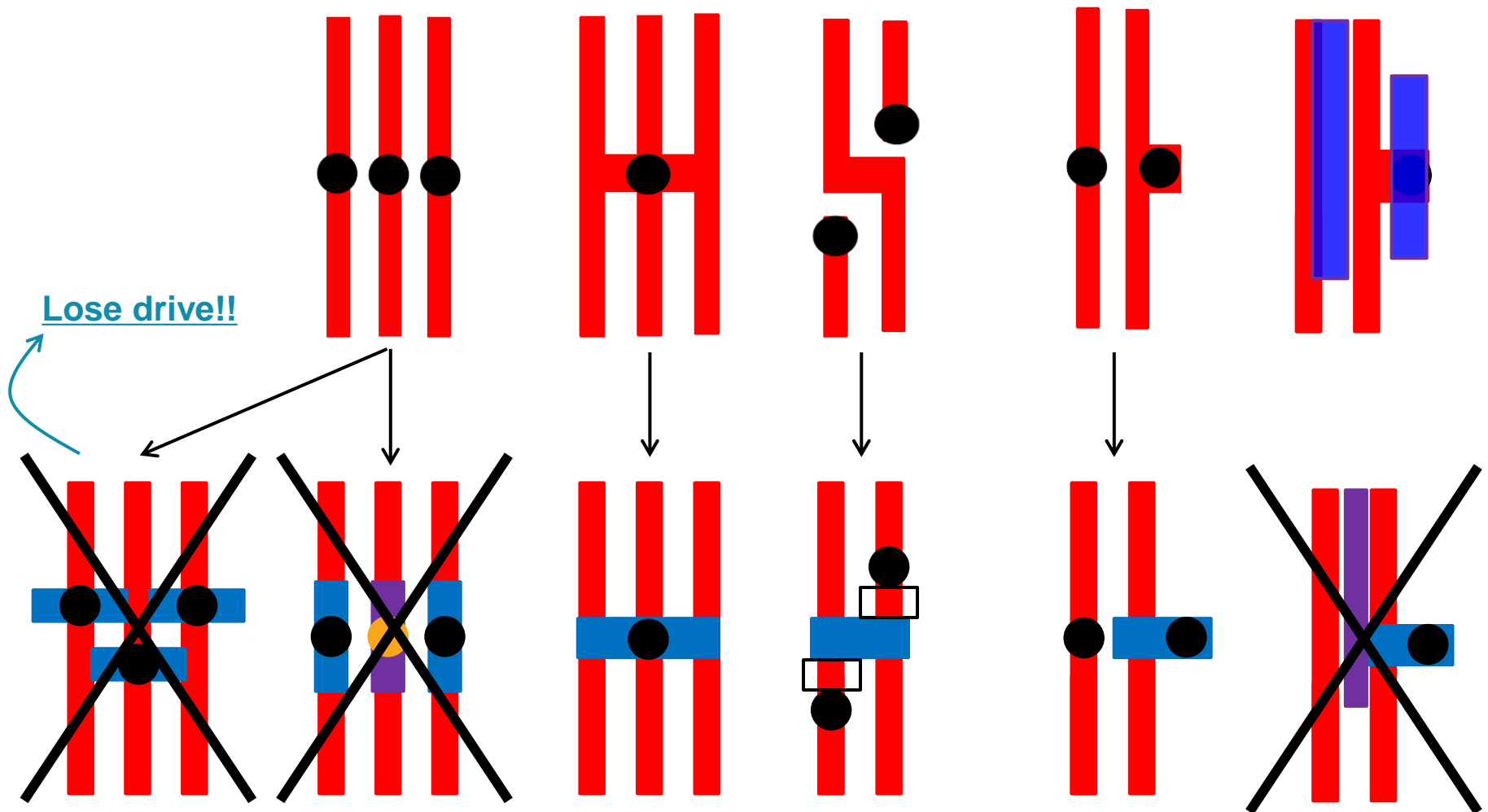


Folded transistor
contact



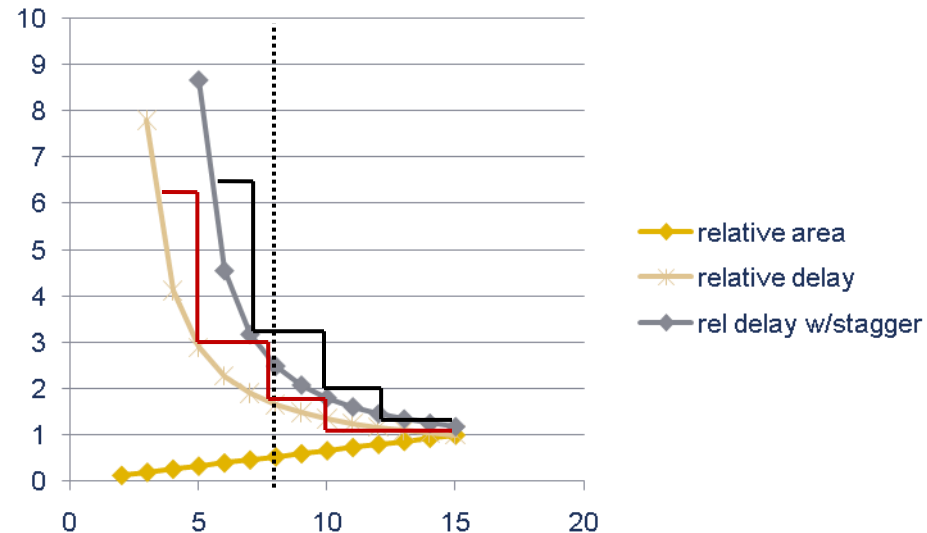
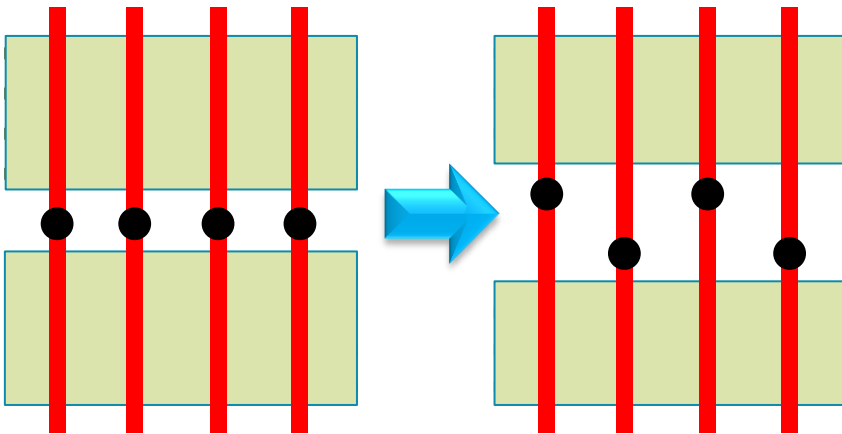
Replacing Lost Cell Constructs with LI

- Primarily replace wrong-way poly w/ LI structures



Staggered Contacts- Why they are bad

- What you get by saving one DP litho step:
 - Low drive == Low performance == Useless Library
 - Real Problem: ratio of load (RC) to transistor area
 - One Example:
 - Staggered 9-track 40% slower than non-staggered
 - Gets worse for smaller track cells (fixed whitespace)



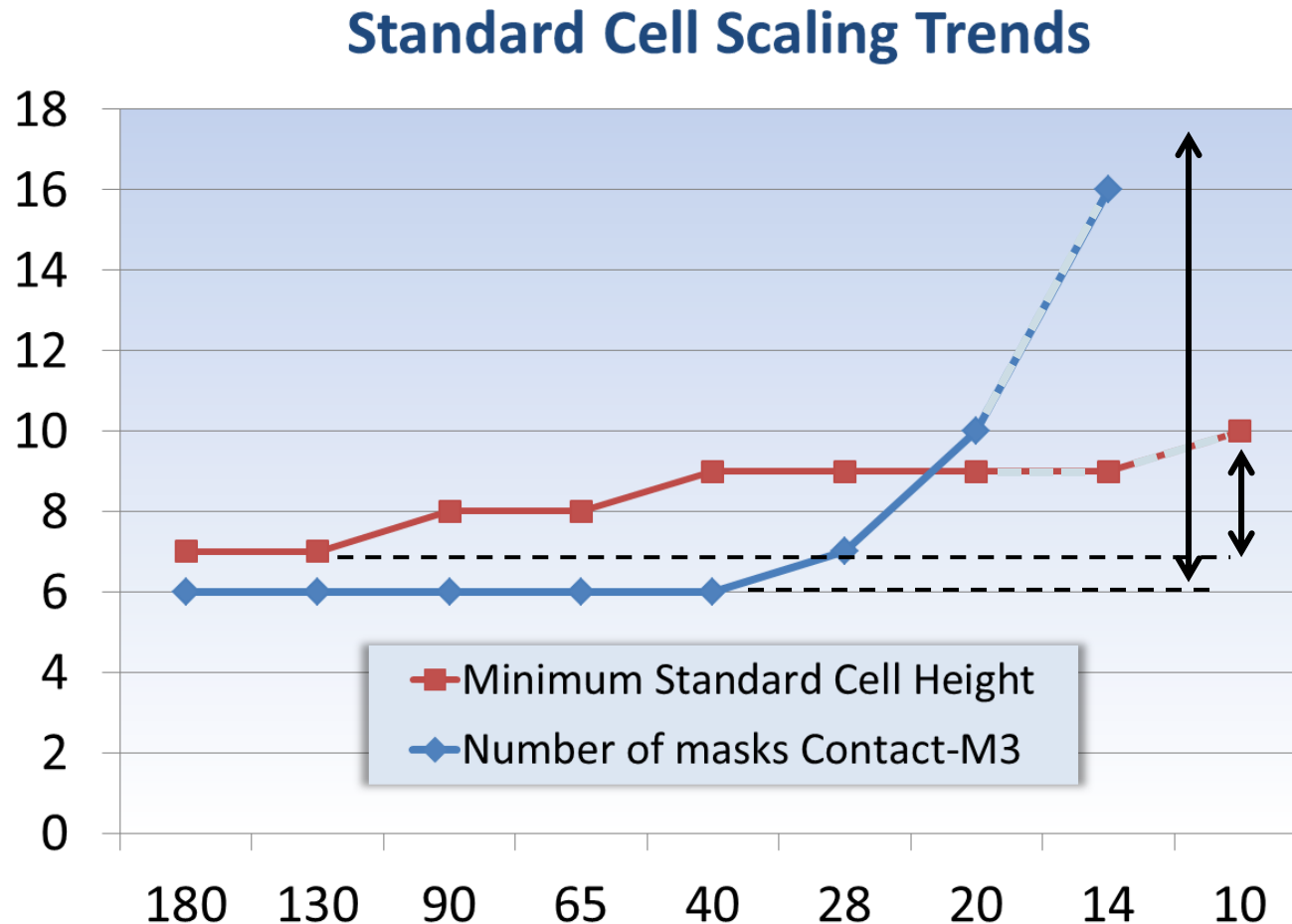
Moving below 110nm Gate Pitch

The “cost-effective” solution is to keep to one mask level by pushing the dipole angle. Source/drain contacts shown in figure

- Tip-to-Tip space inversely proportional to dipole angle
- Bars vs. Holes: More capacitance
- Significant M1 horizontal route blockage
 - Pin quality reduces, horizontal routes blocked
 - Only realistic solution is to add yet another layer: V0
- Number of redundant contacts: 0

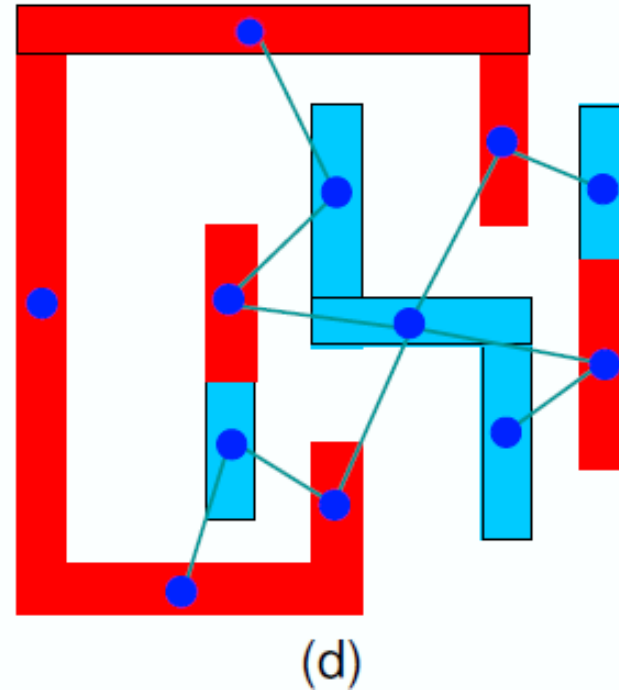
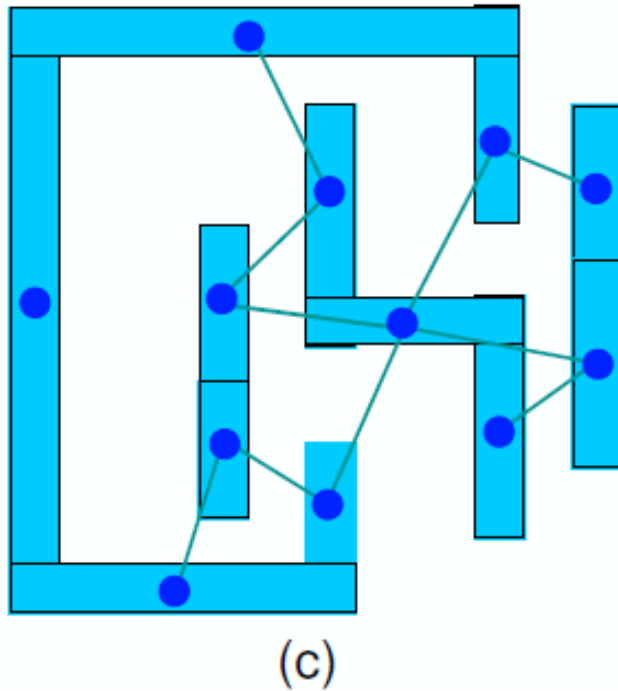


History of Minimum Cell Heights



EUV may be focusing on 20→14 costs but don't forget the 28→20 costs that are already there

M1 DP with LELE



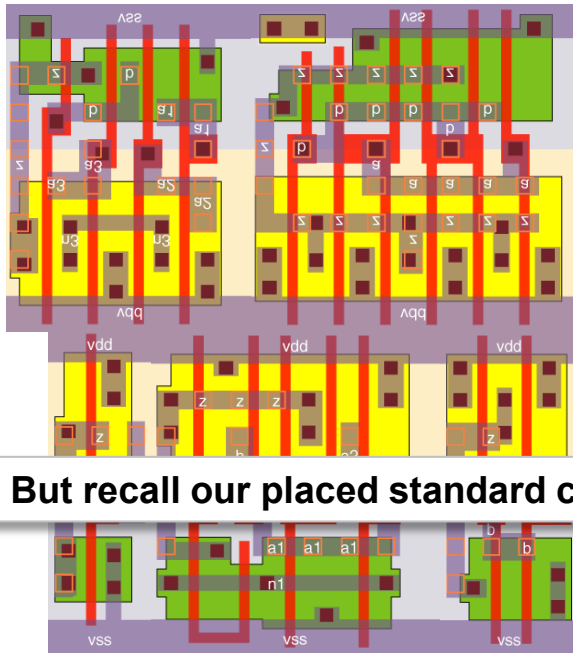
Looks easy enough

A. Kahng, C.-H. Park, X. Xu, and H. Yao, "Layout Decomposition for Double Patterning Lithography", ICCAD 2008, pp 465-474

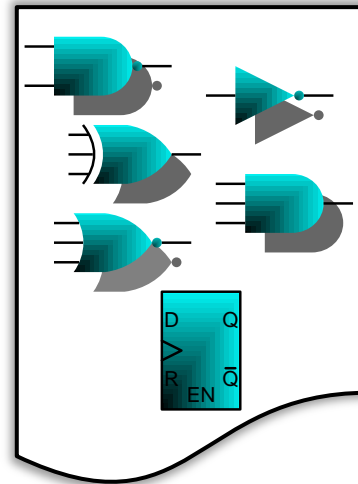
Logic Flow: Place and Route

Required Placements:

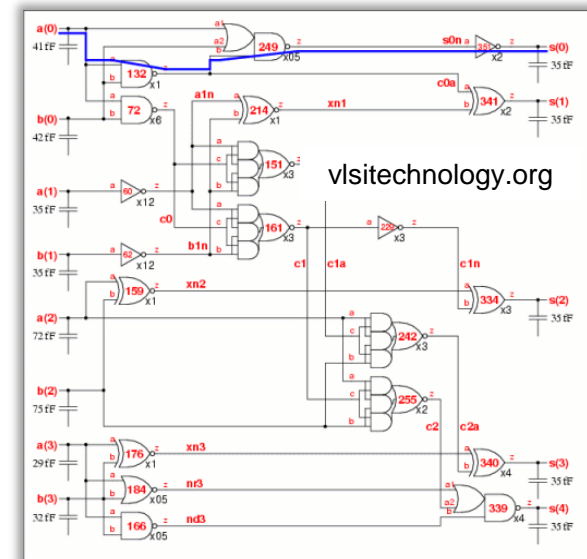
- Vertical and Horizontal Mirroring
- Placement at any gate pitch



But recall our placed standard cells....

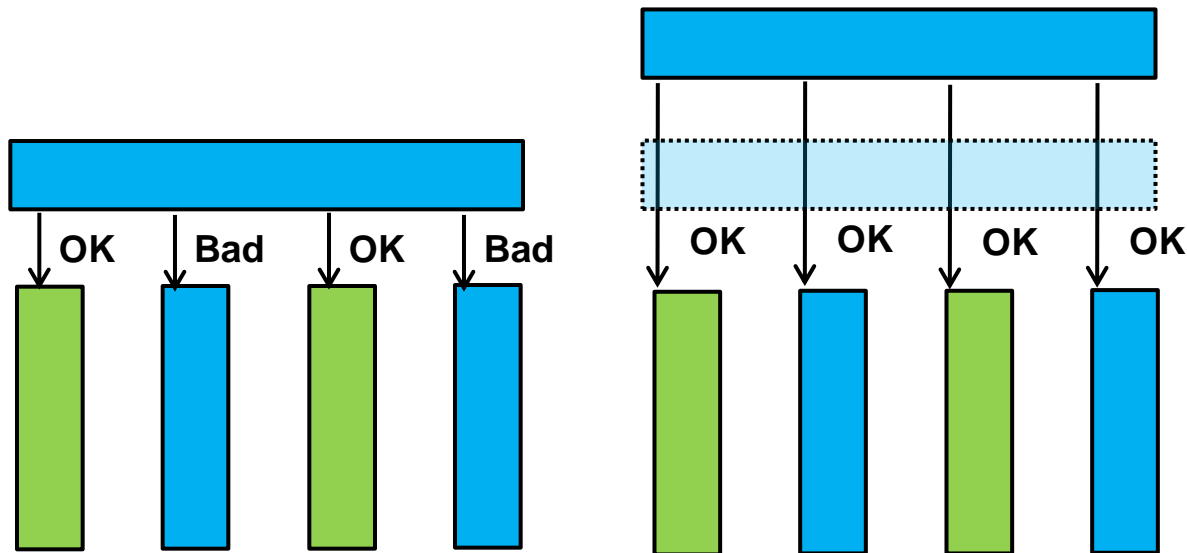


Standard Cell Library



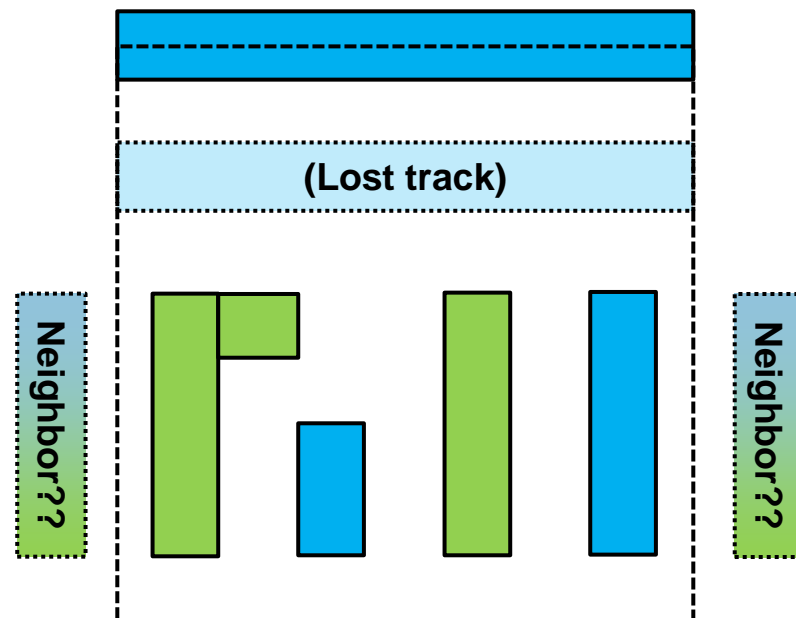
DP: Vertical Trouble for Standard Cells

- Two patterns can be used to put any two objects close together
- Subsequent objects must be spaced at the same-mask spacing
 - Which is much, much bigger (bigger than without double patterning!)
- Classic example: horizontal wires running next to vertical ports
- Two body density not a standard cell problem, 3 body is
- With power rails, can easily lose 4 tracks of internal cell routing!



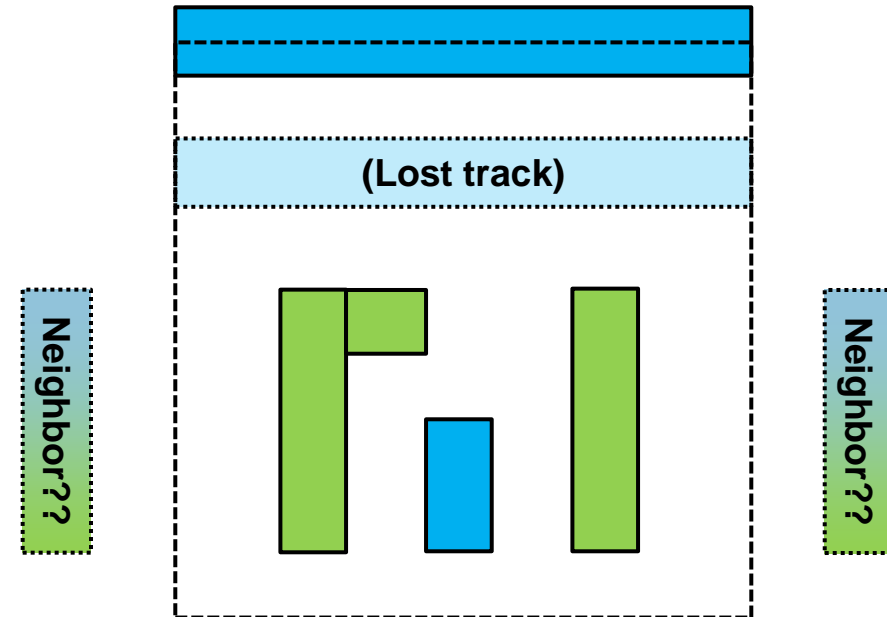
DP: Horizontal Trouble for Standard Cells

- How to meet coloring restrictions for arbitrary left/right cell abutment?



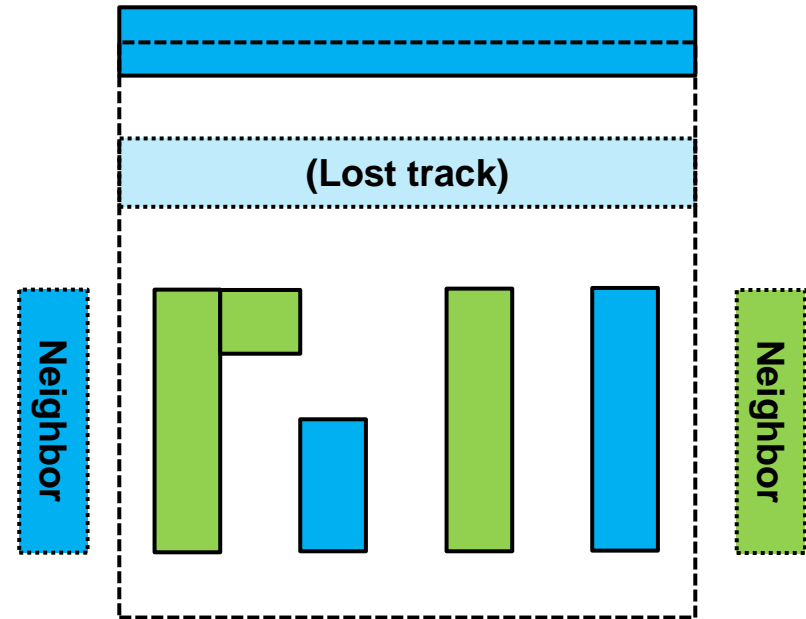
Fix #1: Add in Buffer space

- Guard-banded space at every cell edge is expensive



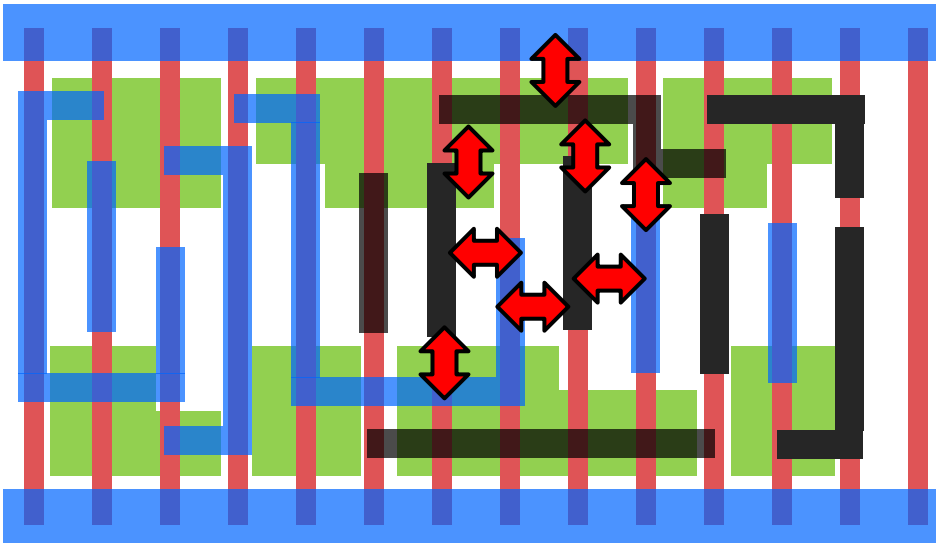
Fix #2: Restrict Placement Options

- Fixing colors is equally expensive
- You can try to save some space by “kicking the can down the road” to the placer, via mirror restrictions.



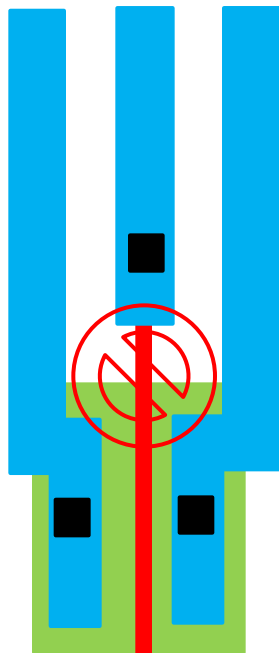
Average Density Vs. Peak Density

- Cases that have only two objects involved historically have had some other way to solve them, so no gain
- E.g., tip-2-side space w/ nothing nearby:
convert to L-shape (get side-2-side space)
- Peak wiring density can not be consistently achieved: Average density is actually closer to the same-mask density, not the different-mask density

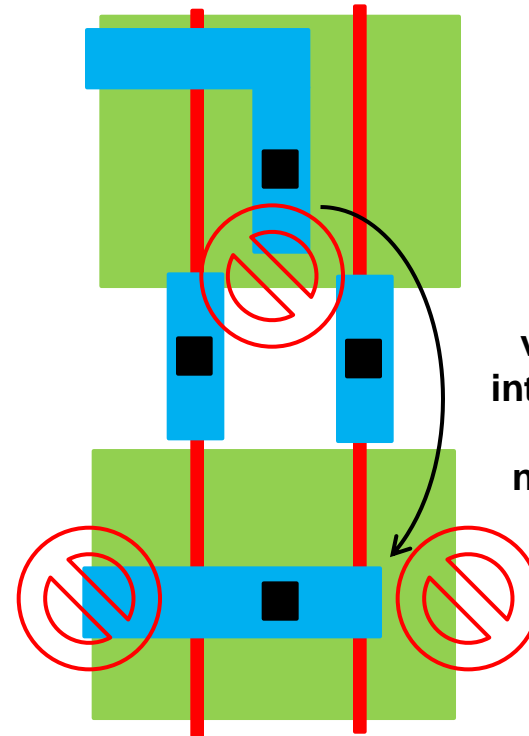
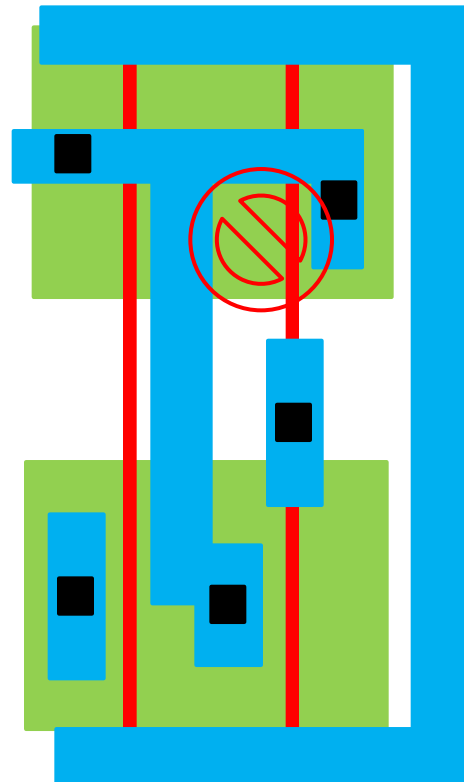


DP: Internal Trouble for Standard Cells

- No small U shapes, no opposing L ‘sandwiches’, etc.
- So several ‘popular’ structures can’t be made
- “Hiding” double patterning makes printable constructs illegal
- Not to mention variability and stitching issues...



No Coloring
Solution

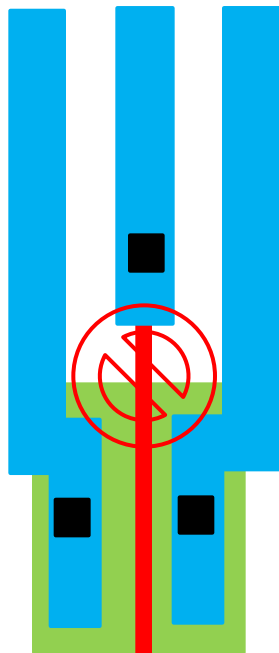


Have to turn
vertical overlaps
into horizontal ones
But that blocks
neighboring sites

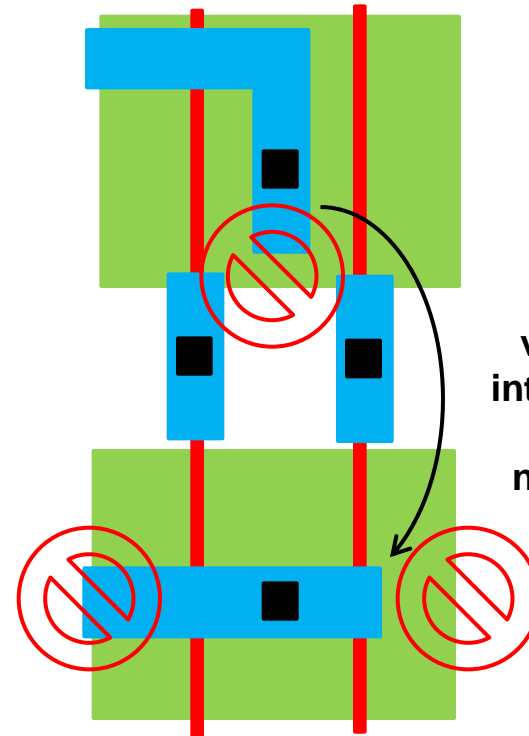
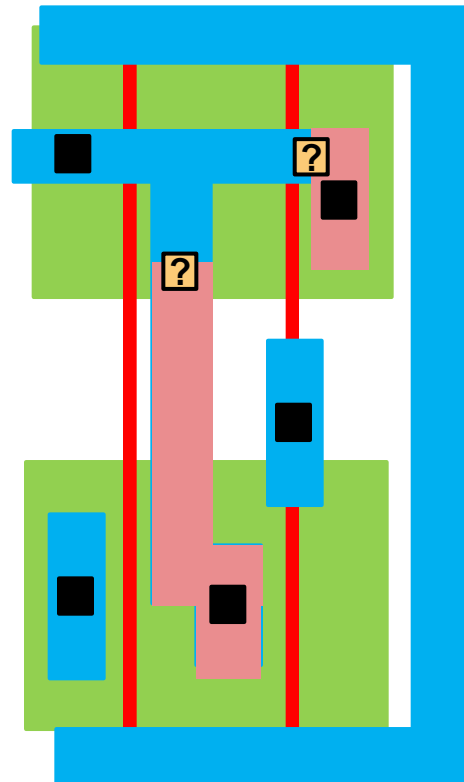
DP: Internal Trouble for Standard Cells

- No small U shapes, no opposing L ‘sandwiches’, etc.
- So several
- “Hiding” double patterning makes printable constructs illegal
- Not to mention variability and stitching issues...

And if we solve problems with color stitching, do we further confine our hit points?



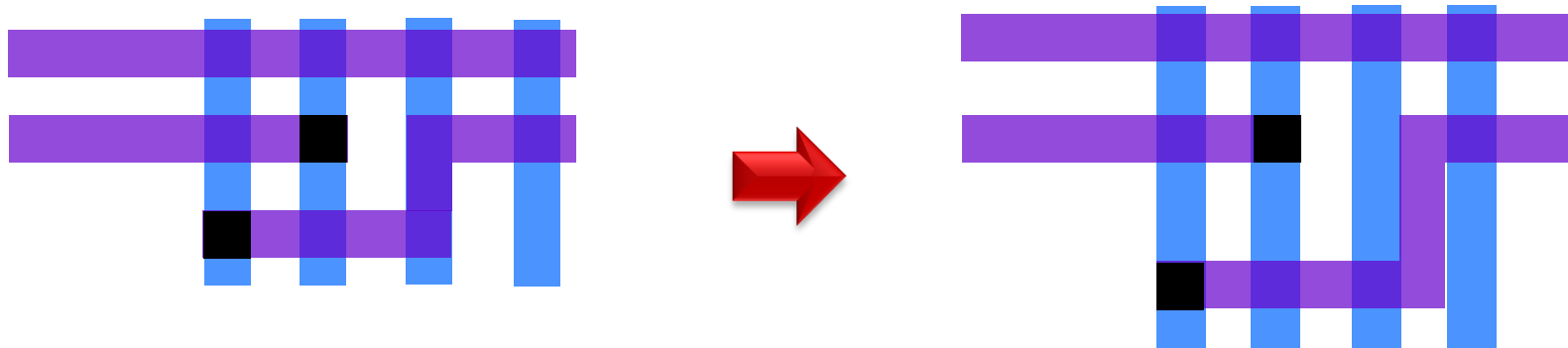
No Coloring Solution



Have to turn vertical overlaps into horizontal ones
But that blocks neighboring sites

Canonical Routing

- 1- and 2- track jogs are very valuable for routing congestion
- But this creates color loops for Double-Patterning
 - Using a hit point blocks other tracks / hit points! (not good)



Liebman and Gutwin, “Quantifying the Design Impact of Double Patterning for the 20nm and 14nm Technology Nodes”, CDNLive! Silicon Valley 2012

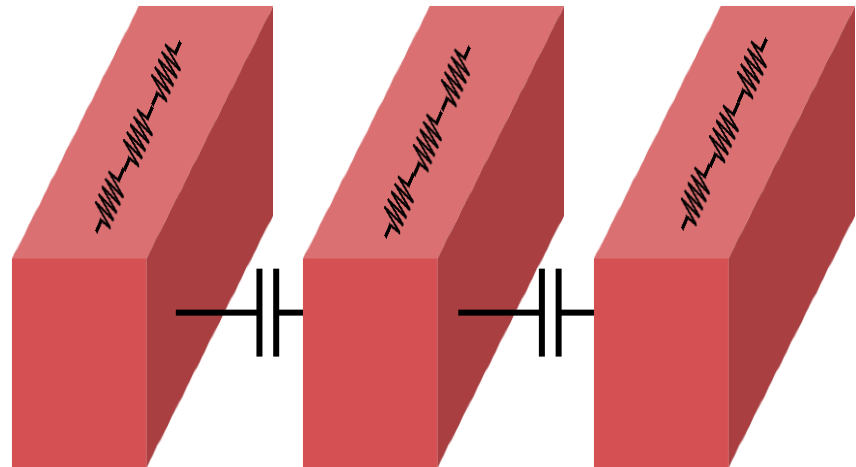
Impact of double patterning in routing layers ~ 5%
(block density 85-90% w/o DP compared to 80-85 with)

Routing Complexity = Cost

- Even at the same pitch, if the rules get more complex, the end product gets larger
- Routing Turn-Around-Time is key
 - 7-10 days run times are common!
 - And this is a batch submission of many parallel runs
 - Seeded by random numbers, slower TAT means fewer shots at the best answer
 - TAT also important to explore possibilities for the microarchitecture

BEOL Signoff

- BEOL assessment happens more than once: Timing, Signal Integrity, Power Delivery, etc.
- Traditionally, 5 corners
 - Typical
 - Worst C
 - Best C
 - Worst RC
 - Best RC



BEOL Signoff

■ With Double Patterning, 15 corners:

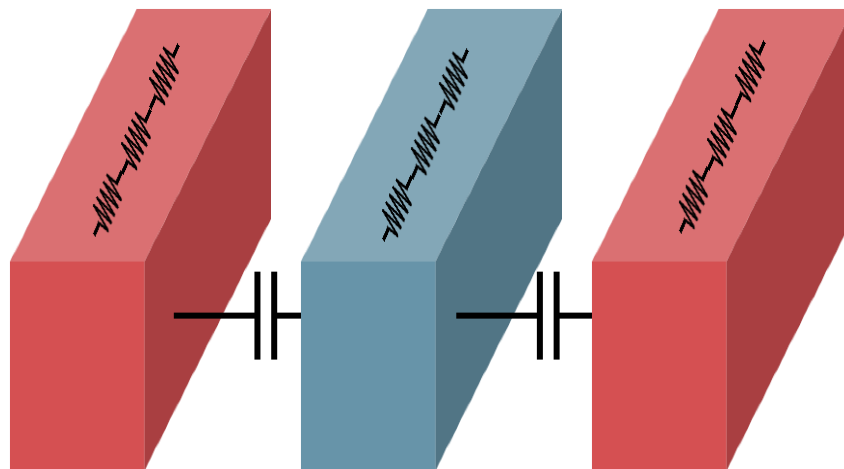
- Typical
- Worst C
- Best C
- Worst RC
- Best RC

X

- Nominal alignment
- Worst Misalignment
- Best Misalignment

> +/- 5%

**Cost Added to Design Infrastructure:
Multi-Valued SPEF + STAR**



Effect of Double Patterning on Wires

- Increases capacitance corners for timing signoff
- Effectively “Increases capacitance”, requiring larger drive cells

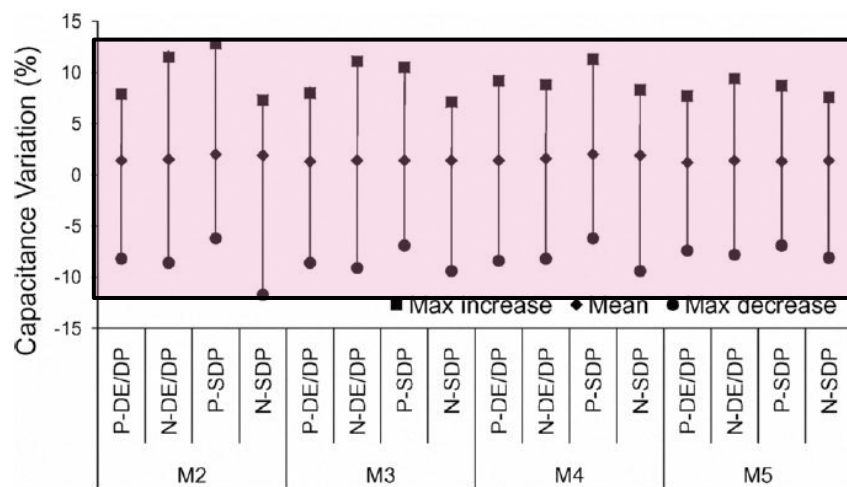
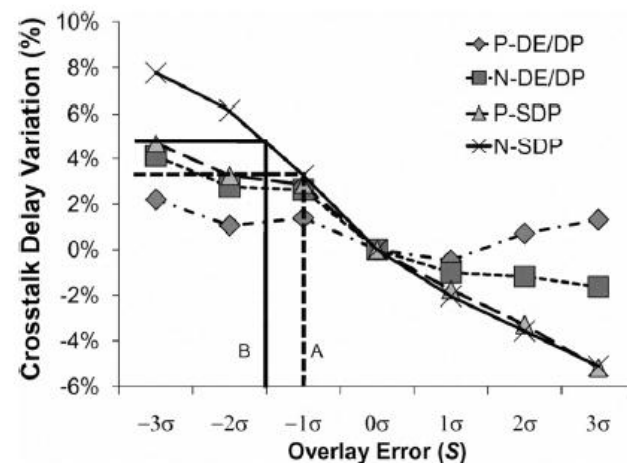


Fig. 9. Capacitance changes (%) of high-capacitance nets ($\geq 2fF$) from 3σ overlay.



K. Jeong, A. B. Kahng, and R. O. Topaloglu, ISQED 2011

Effect of Double Patterning on Wires

Table4: Interconnect parameter extracted from scaled wires. Highest values are in colored cells.

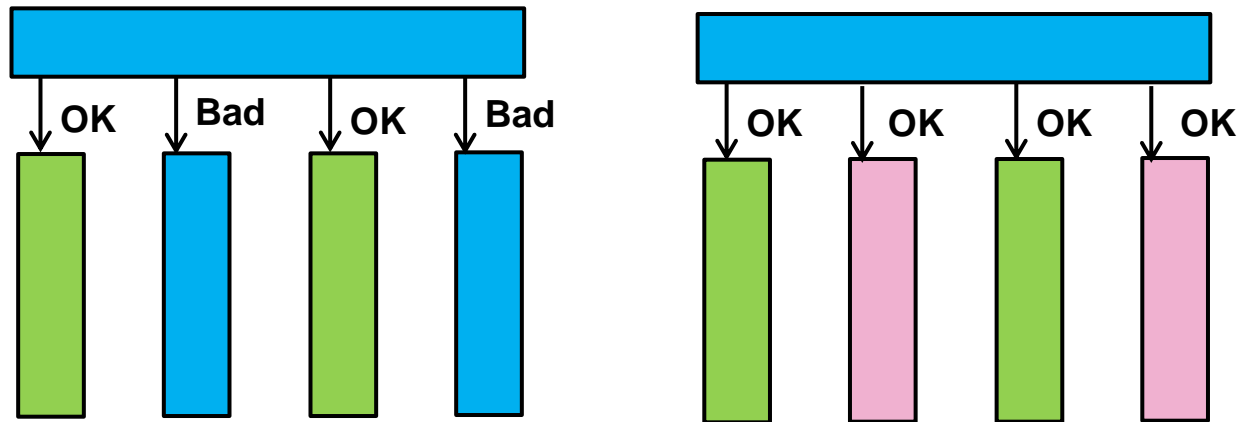
	LELE						SDDP						EUV		
	Wire A			Wire B			Wire A				Wire B			Wire A or B	
Parameter	Nom.	Max.	%	Nom.	Max.	%	Nom.	Max.	%	Nom.	Max.	%	Nom.	Max.	%
R [$\Omega/\mu\text{m}$]	27	31	14%	19	21	14%	18	20	11%	14	16	10%	29	35	20%
C [pF/ μm]	2.5E-4	3.0E-4	22%	2.5E-4	3.1E-4	21%	2.9E-4	2.9E-4	2%	2.8E-4	2.9E-4	2%	3.3E-4	3.5E-4	9%
RC [ps/ μm^2]	6.8E-3	8.2E-3	20%	4.7E-3	5.7E-3	20%	5.0E-3	5.6E-3	12%	4.1E-3	4.6E-3	12%	9.4E-3	1.0E-2	8%
Coupling	Nom.		Max.		%		Nom.		Max.		%		Nom.	Max.	%
C_{AB}/C_{wire}	54%		66%		22%		48%		48%		0%		48%	48%	0%
C_{BA}/C_{wire}	43%		54%		26%		49%		49%		0%		48%	48%	0%
Unbalance															
ΔR_{AB}	88%						55%						0%		
ΔC_{AB}	2%						1.5%						0%		
ΔRC_{AB}	84%						53%						0%		
Δcoupl_{AB}	111%						3%						0%		

M. Stucchi, Z. Tőkei, S. Demuynck and Y.-K. Siew, IMEC, IITC 2012

Summary: Double Patterning

- Minimum pitch (“peak density”) advertised. But actual density answer depends on evaluating cell specifics:
 - How to handle the rails?
 - How to handle left/right cell adjacencies
 - Cells that do not have dense pin configurations will scale better
 - Router inefficiencies
- Once again, must evaluate a full library in order to properly evaluate the average density achievable
- Block implementation is crucial now that DP extends into the routing layers
- Don’t forget signoff / corner modeling implications
- DP costs come in many secondary ways

Triple Patterning: 50% Better?

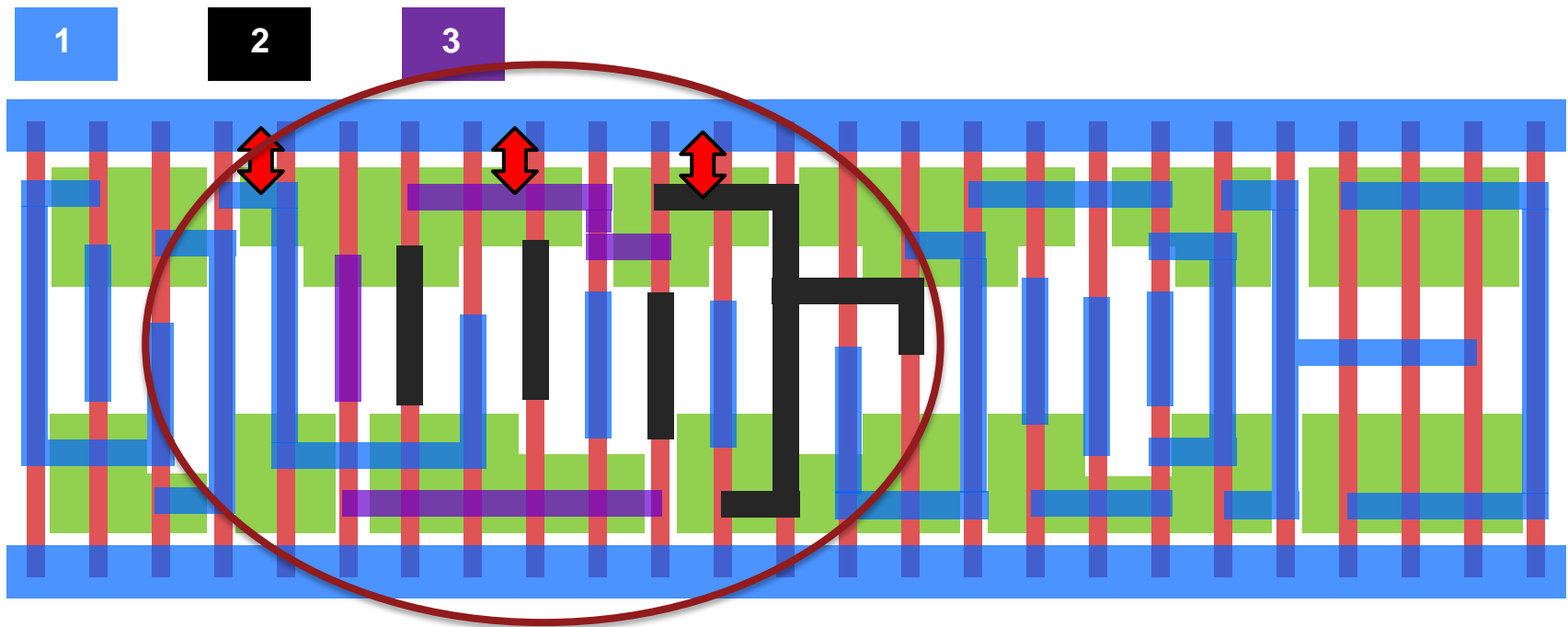


No— average density problem is worse. Router problems are worse.

$$\text{Logic area} \propto N_{\text{MP}} / k_1$$

N_{MP} : Number of Multiple Patterning Steps

Example TP DRC Error



Can a DRC tool tell me if this error is solvable?

Process complexity, error finding and reporting complexity,
large physical span of errors to be fixed
.... this sounds really complicated

Scaling Summary: 28nm to EUV

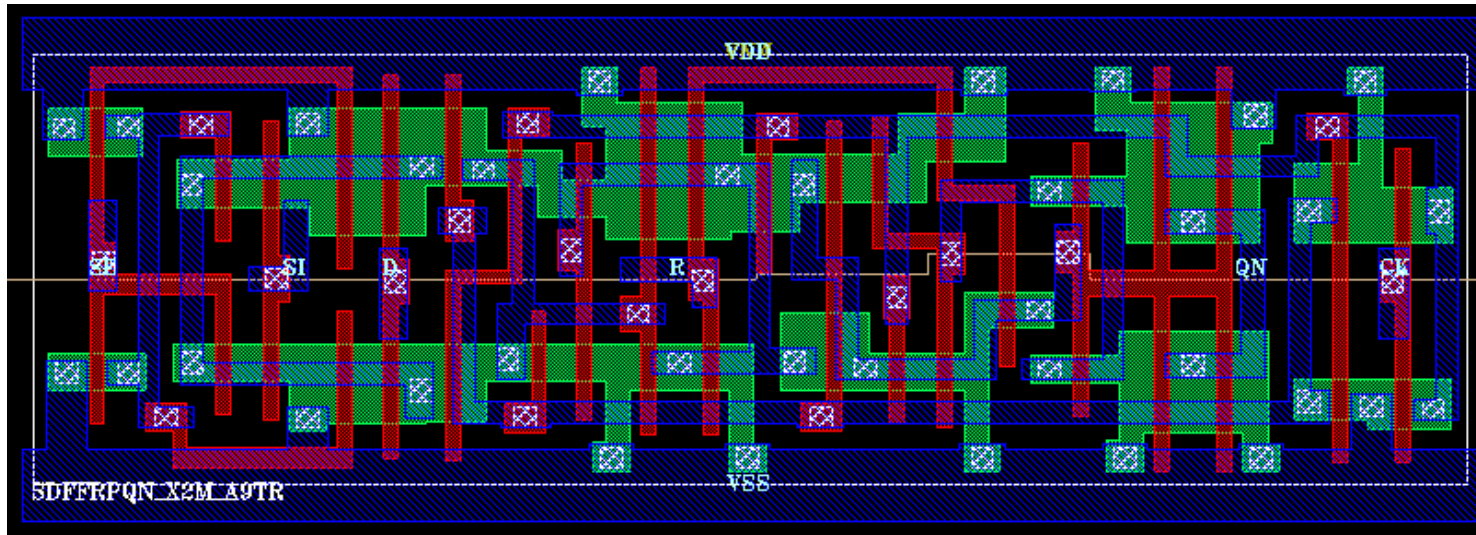
■ Cell Height:	0 - 5%
■ Loss outer channel gate connections	1 - 4%
■ Recapturing of some local routing through LI	0%
■ Multiple Patterning, same-color penalties	4 - 8%
■ Increased use of M2 in complex cells	2 - 4%
■ Increased complexity in routing layers	4 - 10%
■ Misc: (DP signoff, via restrictions, etc.)	1 - 3%

28 to EUV:

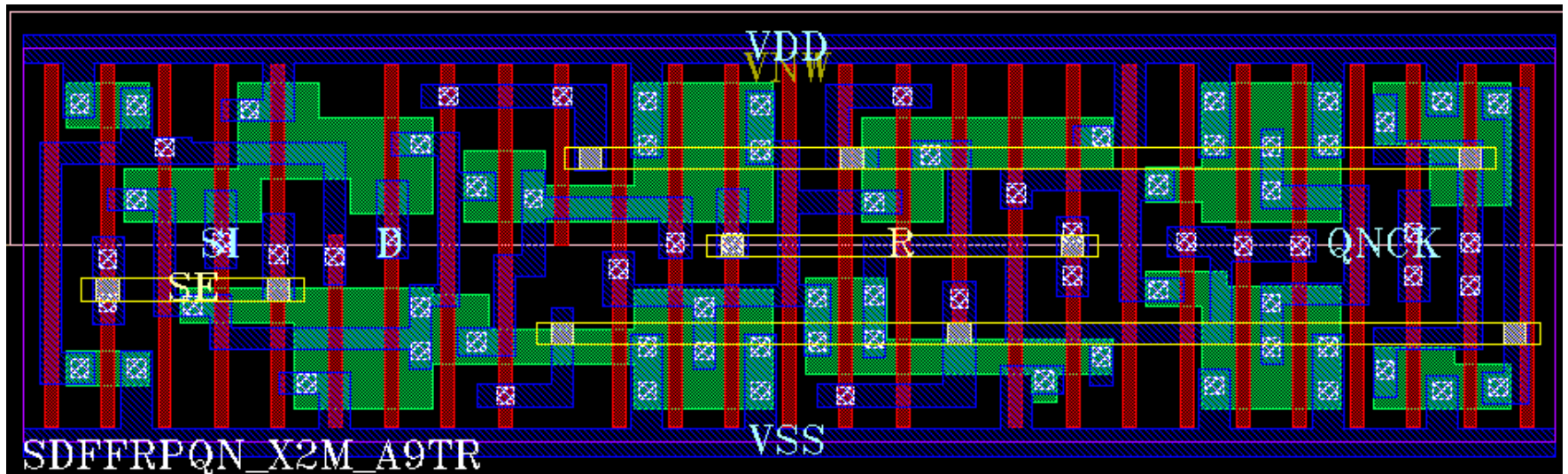
Another half node of area scaling lost!

65nm to EUV: The Lost Node

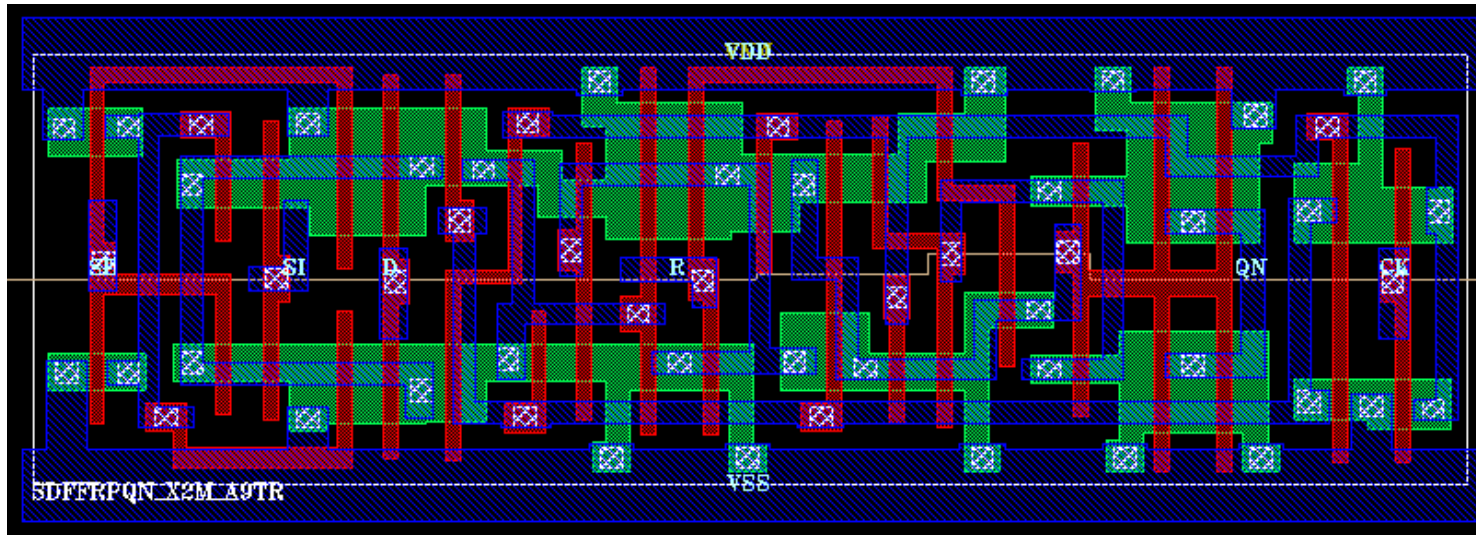
Can EUV turn back the k_1 lock?



32nm



Can EUV turn back the k_1 lock?



10nm (?)

With Fins, probably not on active

- Non-rectangular shapes and epitaxial growth might not go together
- Fin R vs contact R

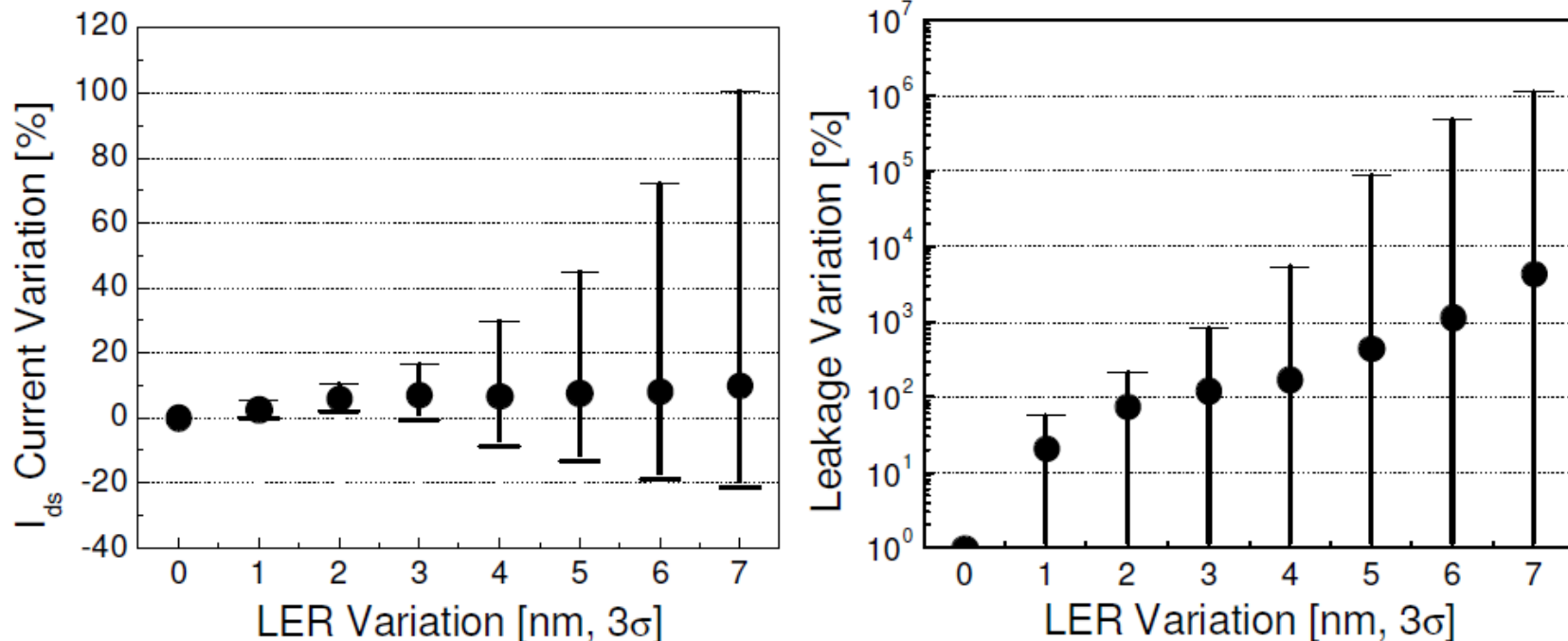
With metal gate, probably tough to give up the uniformity (CMP, liner fill, etc.)

- but maybe a gate tab for contact offset would be workable?

LER would be a key start to either conversation!

LER and Design Margins

Design flows care about the error bars



(a) LER impact on I_{on}

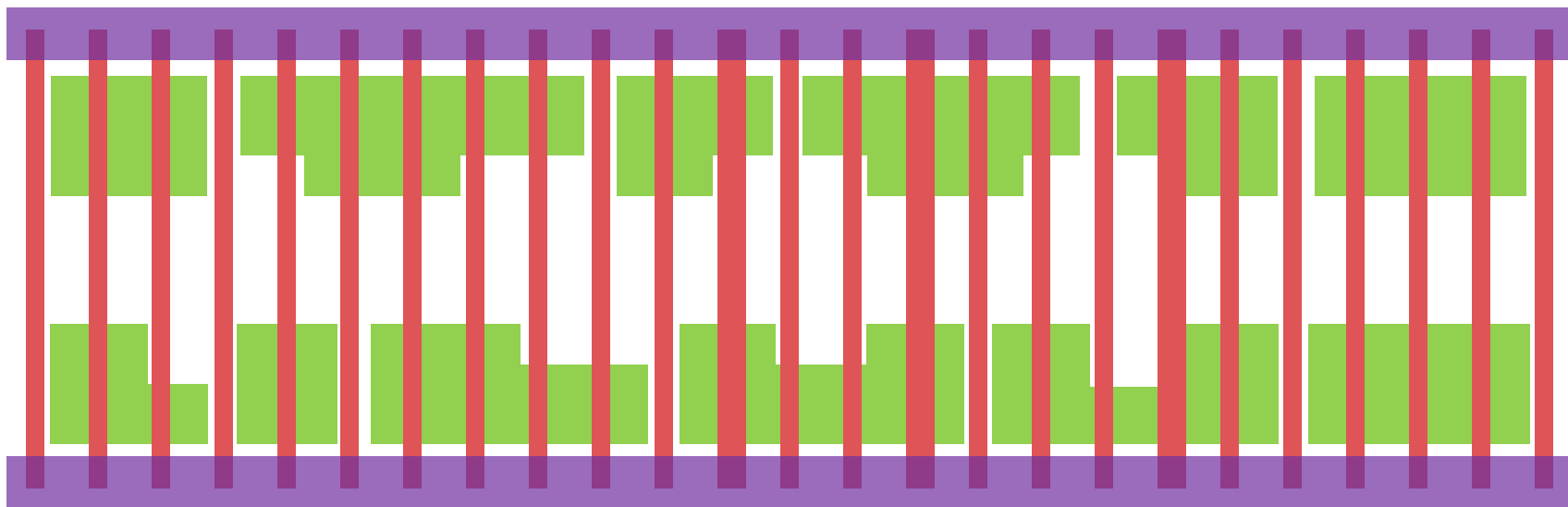
(b) LER impact on I_{off}

Figure 2: LER impact on I_{on} and I_{off}

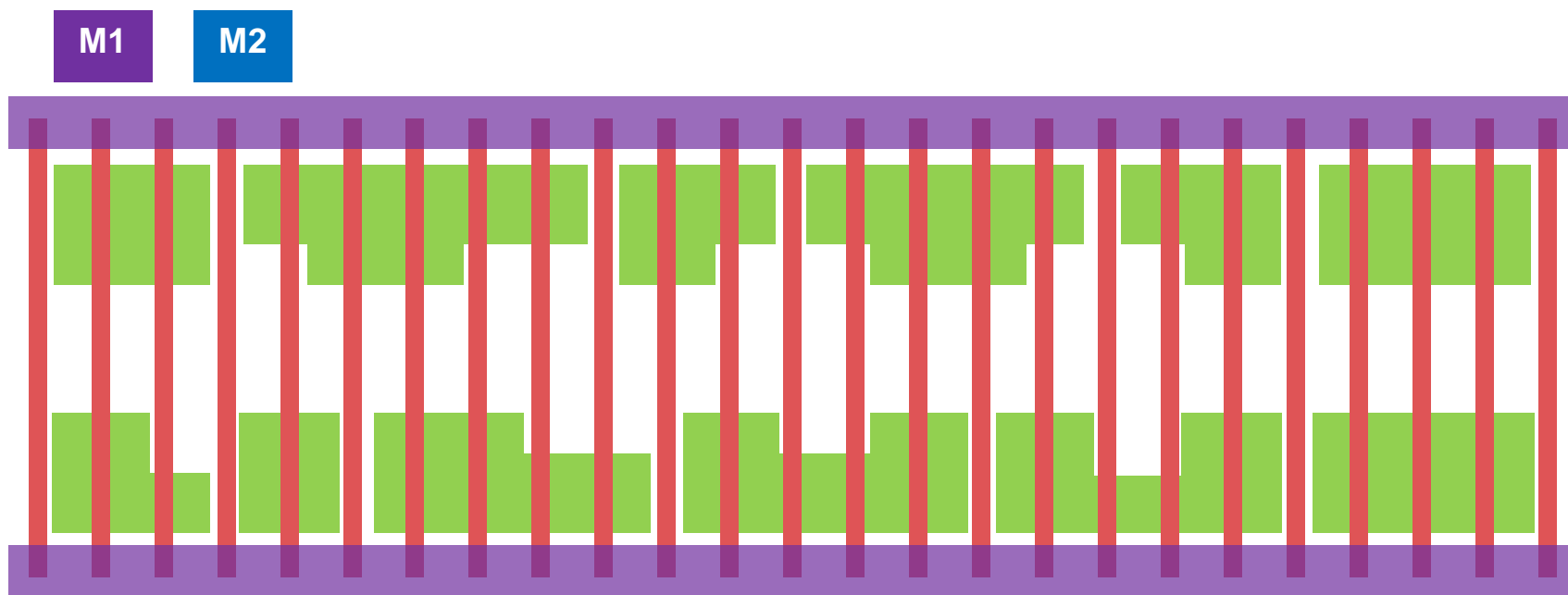
Y. Ban and J.-S. Yang, DAC 2011

EUV Opportunity: Gate Sizing

- The number of gate length options is shrinking with each technology node
 - Reduces ability to optimize power and performance
- Gate length transition Pressure becoming significant
- Not allowed for many generations: Local gate sizing
- Perhaps EUV can re-open possibility of intra-cell transistor sizing



Pressure on Active space, vs. gate pitch



**Increasing active space to 1nm more than will fit in a gate pitch
=> 20% area penalty!**

EUV Opportunity: Contact

- Can we revert from slots to holes?
- Generally, trading off more R for less C is beneficial
- Cost benefit of redundancy is not well quantified, but probably ends up being a positive result
- If we can get back to contact spaces = gate spaces, we can go back to some higher density layouts

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

56

57

58

59

60

61

62

63

64

65

66

67

68

69

70

71

72

73

74

75

76

77

78

79

80

81

82

83

84

85

86

87

88

89

90

91

92

93

94

95

96

97

98

99

100

101

102

103

104

105

106

107

108

109

110

111

112

113

114

115

116

117

118

119

120

121

122

123

124

125

126

127

128

129

130

131

132

133

134

135

136

137

138

139

140

141

142

143

144

145

146

147

148

149

150

151

152

153

154

155

156

157

158

159

160

161

162

163

164

165

166

167

168

169

170

171

172

173

174

175

176

177

178

179

180

181

182

183

184

185

186

187

188

189

190

191

192

193

194

195

196

197

198

199

200

201

202

203

204

205

206

207

208

209

210

211

212

213

214

215

216

217

218

219

220

221

222

223

224

225

226

227

228

229

230

231

232

233

234

235

236

237

238

239

240

241

242

243

244

245

246

247

248

249

250

251

252

253

254

255

256

257

258

259

260

261

262

263

264

265

266

267

268

269

270

271

272

273

274

275

276

277

278

279

280

281

282

283

284

285

286

287

288

289

290

291

292

293

294

295

296

297

298

299

300

301

302

303

304

305

306

307

308

309

310

311

312

313

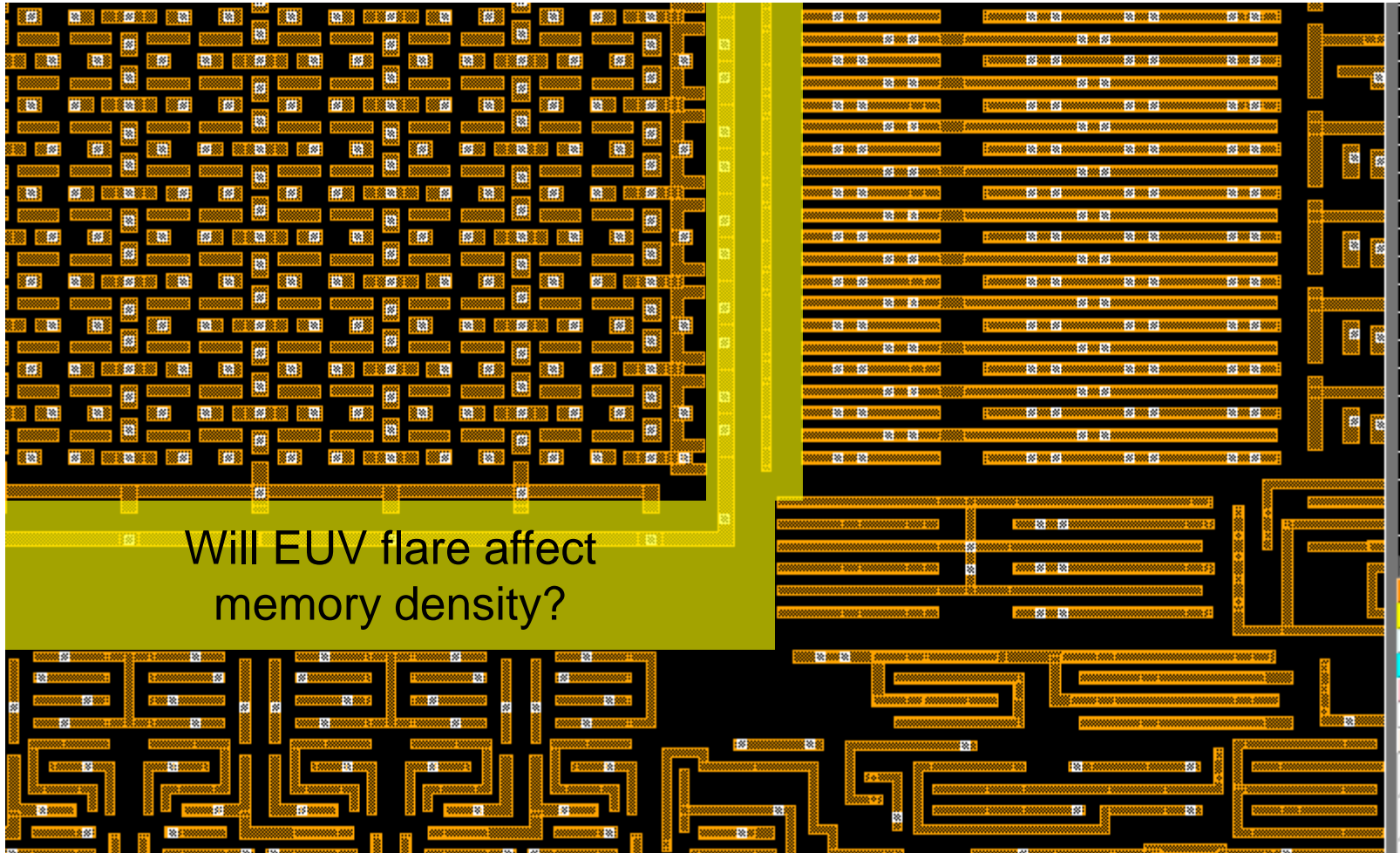
314

315

316

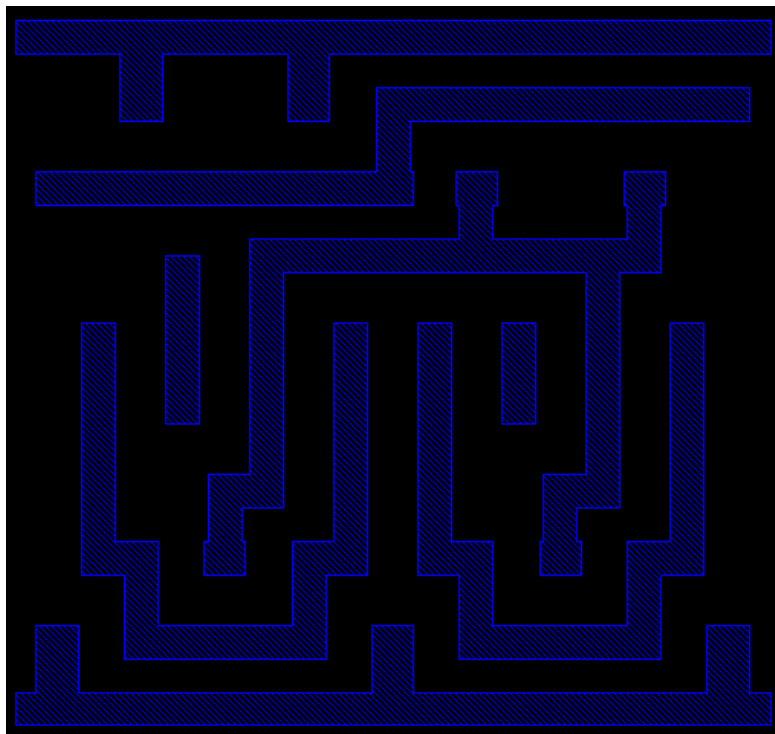
317

SRAM Array-Periphery Transition



EUV and M1

- Yes, this is the key opportunity
- With modern technologies, M1 routing is not allowed because the rules are too complex. If the router can get involved again in M1, will see significant area savings



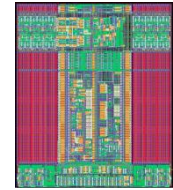
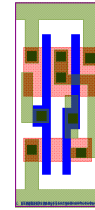
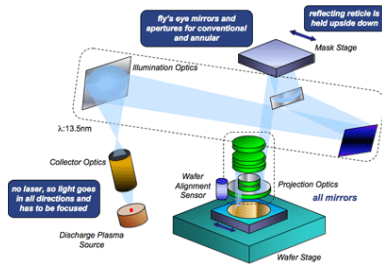
EUV and Mx/Vx

- It's not just about who can do the minimum pitch in the minimum cost
 - Line ends, with vias
 - At-pitch contacting
 - Router complexity effects
 - 5% DP penalty will surely increase with TP/QP
 - And, once again, look at effects such as capacitance variation on actual implemented blocks.

Summary

- Don't forget history: The Lost Node may add to the EUV value proposition
- To evaluate any technology, one must examine all design rules and how they interact, with all physical IP instances, and the implementation flow
 - The answer is always “it depends” (on circuit, targets, etc.)
 - One might be tempted to infer too much from a single NAND gate, but the real answer is complex and statistical
 - Real answers require routing and block level metrics (and that investment also likely beneficial to the EUV cause)
- Detailed Design-Technology Co-Optimization can help you tune your solution for optimum benefit
 - To borrow from U.S. Sen. Everett Dirksen: “a nanometer here, a nanometer there, pretty soon you are talking about real money”





Thank You!

Have a productive conference-- Designers are rooting for you!
(don't forget about us)

And thank you to contributing co-workers:

Rob Aitken

David Pietromonaco

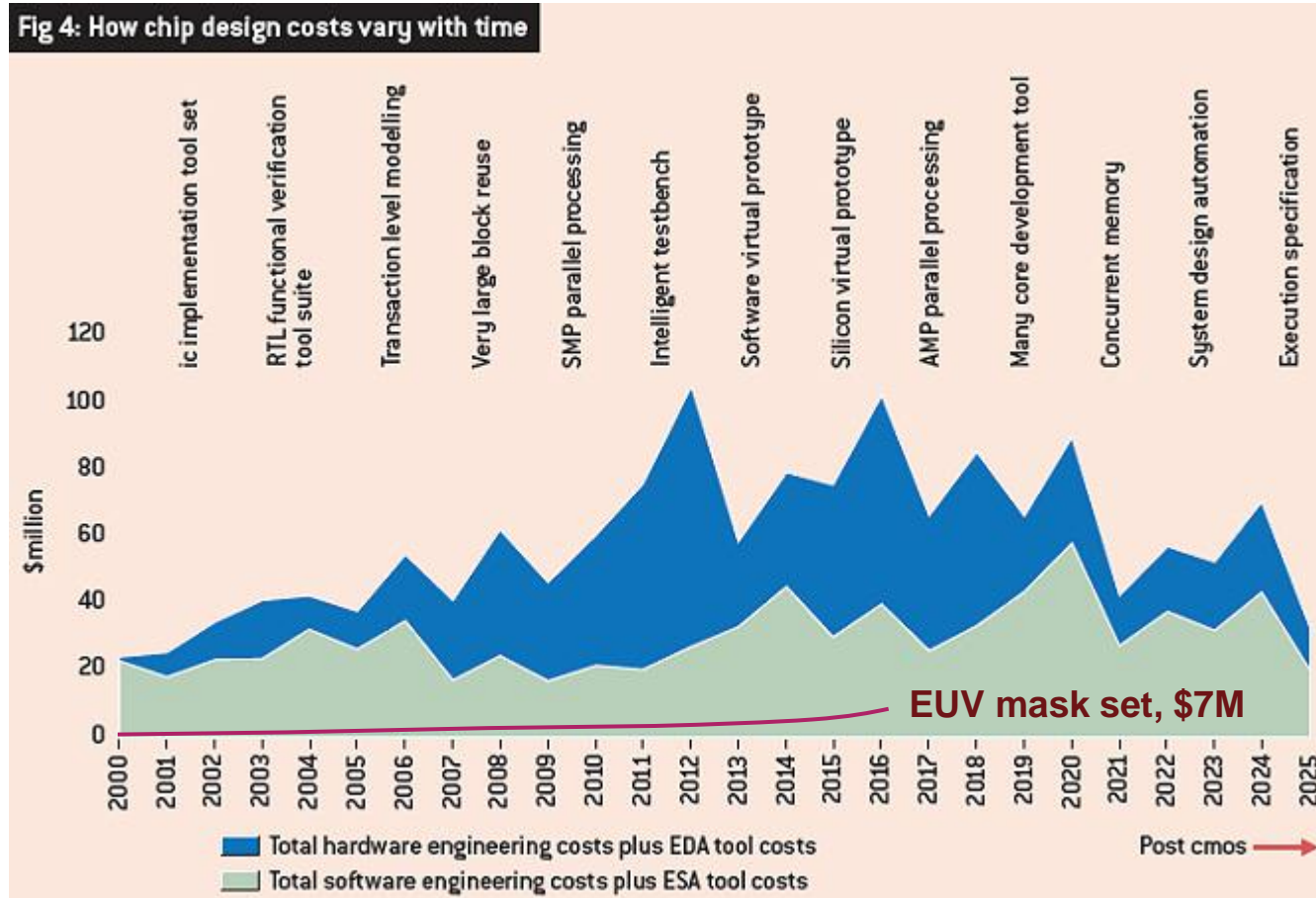
Brian Cline



Additional Information

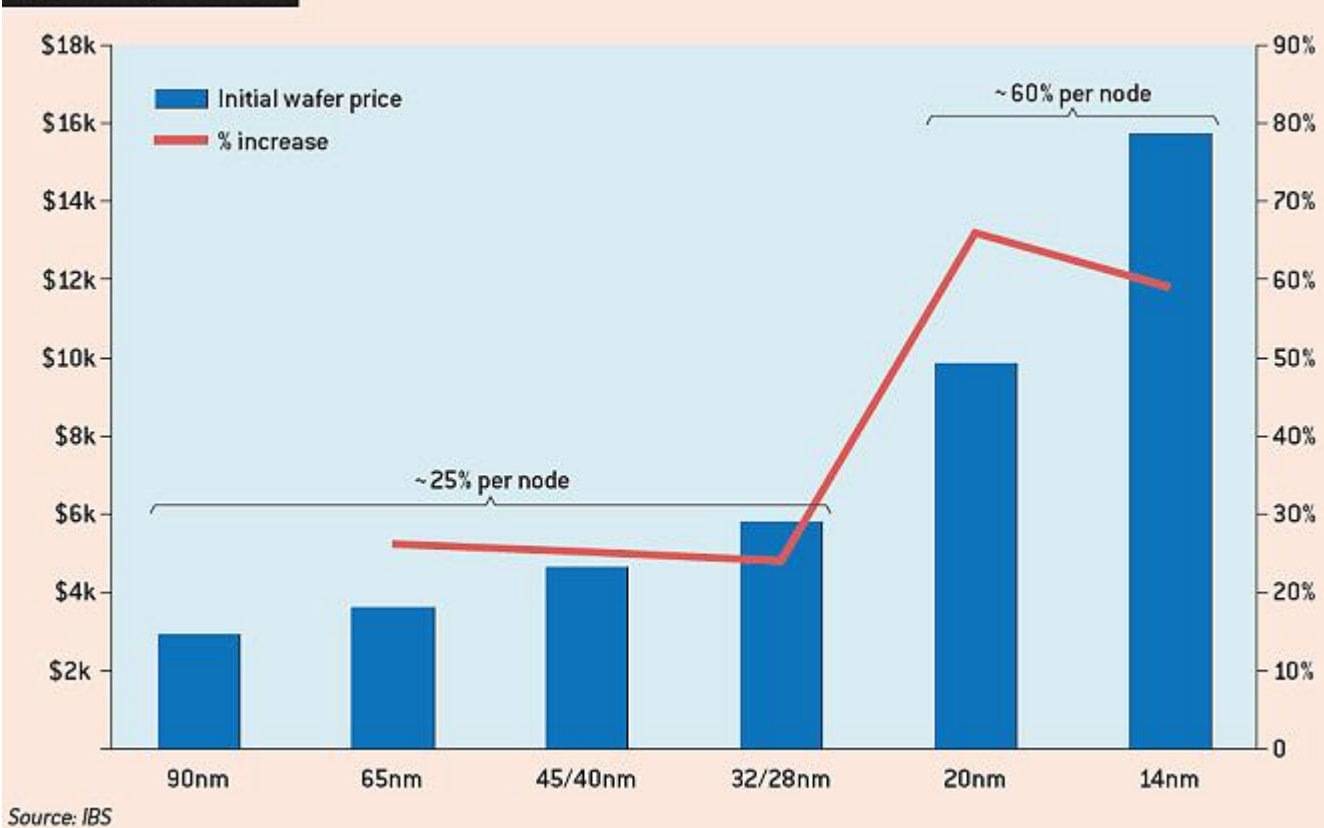


Design Cost



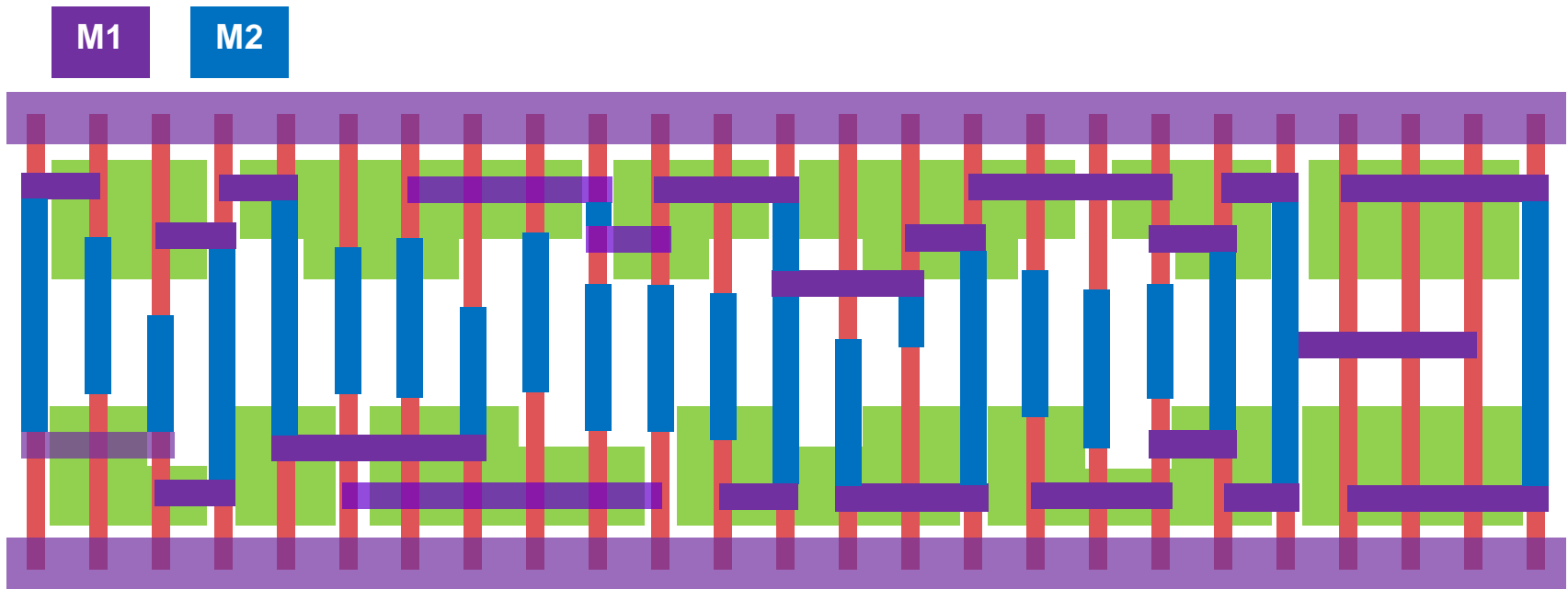
Chris Edwards, "The economics of chip manufacture on advanced technologies", newelectronics.co.uk

Fig 2: Wafer price trend



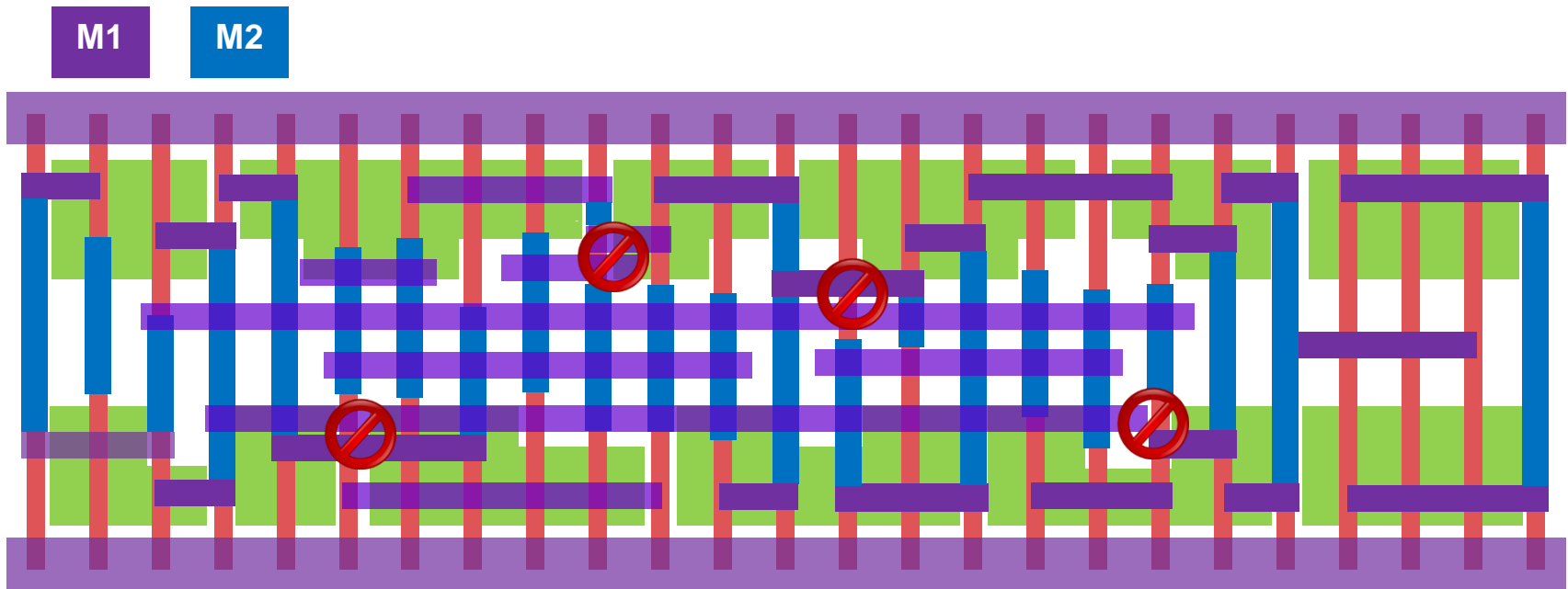
Chris Edwards, "The economics of chip manufacture on advanced technologies", newelectronics.co.uk

What About Uni-Directional Metal?



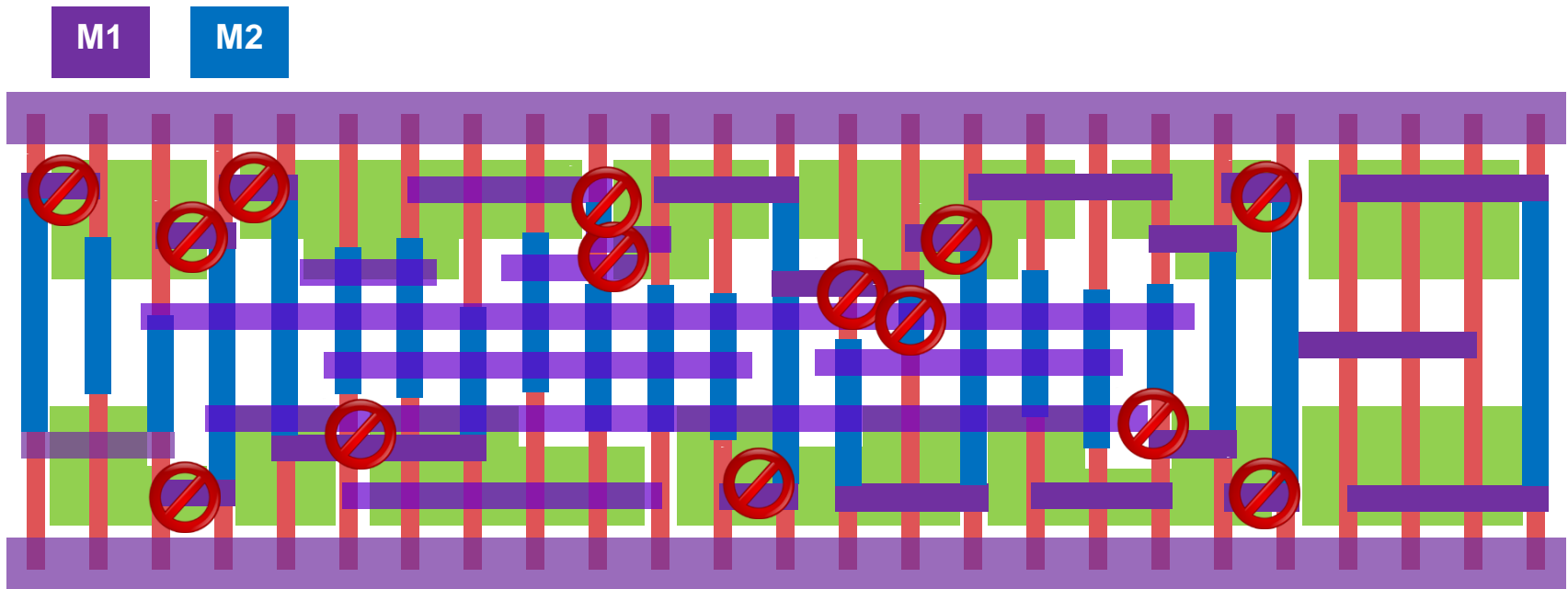
Why work that hard?
This looks easier!

What About Uni-Directional Metal?



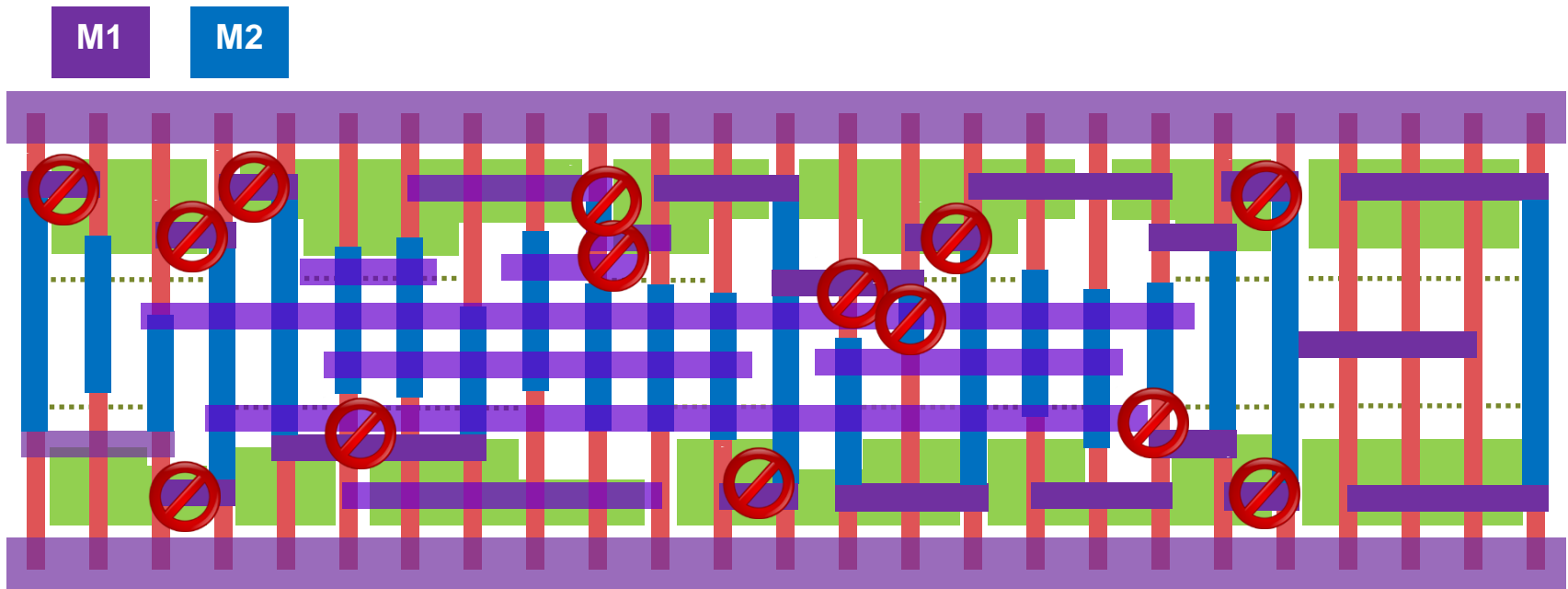
1. Complex cells already use M2, would need to expand

What About Uni-Directional Metal?



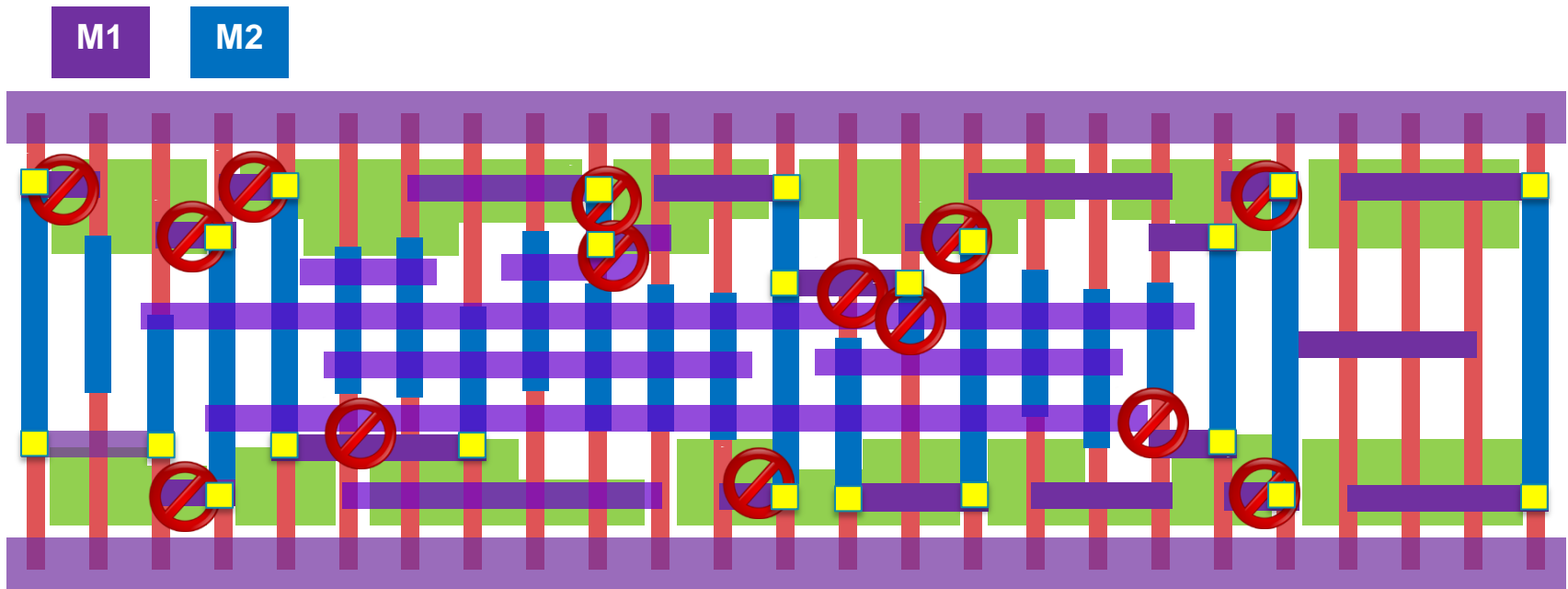
1. Complex cells already use M2, would need to expand
2. Minimum metal areas have not been scaling, now very costly

What About Uni-Directional Metal?



1. Complex cells already use M2, would need to expand
2. Minimum metal areas have not been scaling, now very costly
3. Try contacting adjacent gates with these minimum metal shapes
4. The only way to make room for the metal is to shrink the transistors, and the performance cost = area and power

What About Uni-Directional Metal?



1. Complex cells already use M2, would need to expand
2. Minimum metal areas have not been scaling, now very costly
3. Try contacting adjacent gates with these minimum metal shapes
4. The only way to make room for the metal is to shrink the transistors, and the performance cost = area and power
5. Significant increase in the number of vias